UNED Escuela Internacional de Doctorado EIDUNED

PhD Thesis

2021

THE OBSERVATIONAL REPRESENTATION FRAMEWORK AND ITS IMPLICATIONS IN DOCUMENT SIMILARITY, FEATURE AGGREGATION AND RANKING FUSION

FERNANDO GINER MARTÍNEZ

DOCTORAL PROGRAMME IN
INTELLIGENT SYSTEMS

ENRIQUE AMIGÓ CABRERA

To my parents

Acknowledgements

First and foremost I want to thank my advisor Dr. Enrique Amigó. It has been an honor to be his Ph.D. student. He has taught me how good scientific research is done. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example he has provided as a successful man scientific and professor.

I would like to express my sincere gratitude to Dr. Felisa Verdejo, for allowing me to undertake this work. Her collaborations, continuous guidance, advice, effort and suggestions throughout the research made the methodology of this thesis comprehensive. I appreciate her continuous encouragement.

My gratitude is also due to Dr. Julio Gonzalo, for his extraordinary collaborations and support to the whole process of this research project. I was immensely benefited from his guidance throughout the various stages of this study.

I would also like to thank to all my co-authors Dr. Damiano Spina, Dr. Stefano Mizzaro, Dr. Jorge Carrillo-de-Albornoz and Dr. Tamara Martín the fruitful collaboration with them has been of significant benefit for this dissertation.

The members of the UNED group in Natural Language Processing and Information Retrieval have contributed immensely to my personal and professional time at UNED. The group has been a source of friendships as well as good advice and collaboration. I am especially grateful to Dr. Victor Fresno and Dr. Ana García-Serrano.

I would also like to thank Dr. German Rigau and Dr. Miriam Fernández who extended his support as panel members of the Doctoral Consortium, giving me useful guidelines in the development of this thesis.

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Abstract

Programa de Doctorado en Sistemas Inteligentes
Escuela Internacional de Doctorado EIDUNED

Doctor of Philosophy in Computer Science

The Observational Representation Framework and its Implication in Document Similarity, Feature Aggregation and Ranking Fusion

by Fernando Giner Martínez

Document representation is a core issue in information access tasks. Representing documents requires managing features in terms of three aspects: weighting, redundancy and scaling (i.e., quantitative vs. discrete features). In supervised scenarios, this is done by maximizing effectiveness over specific tasks and training data. However, in this thesis, we focus on non-supervised scenarios, in which document representation is guided by how features are distributed throughout a document collection. Based on an analysis of the literature, we claim in this thesis that traditional representation approaches are not able to capture weighting, redundancy and quantitativity simultaneously. In this thesis, we present the Observational Representation Framework (ORF), which overcomes this limitation. The ORF integrates aspects of representation models based on vector spaces, feature sets and information theory. In addition, we explore the theoretical and practical implications of the ORF in three ways. In the first study, we exploit ORF as a formal framework for document similarity. In this study, we identify the strengths and weaknesses of existing similarity functions based on metric spaces (cosine distance, Euclidean distance, etc.), feature sets (Jaccard distance, Dice distance, etc.) and information theory (pointwise mutual information (PMI), Lin's similarity, conditional probability, etc.). To overcome the limitations observed in this analysis, we define the *Information Contrast Model* (ICM), which is a parametrized generalization of the PMI. In the second study, we empirically check the ability of the ORF to integrate heterogeneous features (i.e., features with discrete and continuous values) without requiring supervision. We perform experiments in the context of message clustering for online reputation management. Finally, in the third study, we analyse the ORF as a formal basis for ranking fusion. Our formal analysis shows that the ORF can accommodate different ranking fusion algorithms depending on the assumptions adopted, such as averaging schemes and the Borda, Copeland and Unanimous Improvement Ratio (UIR) algorithms. Our experiments on six ranking fusion datasets shed light on which aspects of the scenarios at hand determine the suitability of different assumptions and ranking fusion algorithms.

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Resumen

Programa de Doctorado en Sistemas Inteligentes
Escuela Internacional de Doctorado EIDUNED

Doctor en Informática

El Marco de Representación Observacional y su Implicación en Similaridad de Documentos, Agregación de Características y Fusión de Rankings.

por Fernando Giner Martínez

La representación de documentos es un elemento clave en tareas de acceso a la información. Representar documentos requiere gestionar características en términos de tres aspectos: ponderación, redundancia y escalado (i.e., características cuantitativas vs. discretas). En escenarios supervisados, éstos se abordan maximizando la efectividad sobre tareas específicas y datos de entrenamiento. Sin embargo, en esta tesis, nos centramos en escenarios no supervisados, donde la representación de documentos está dirigido por la distribución de los rasgos en una colección de documentos. Basado en un análisis de la literatura, afirmamos que las aproximaciones de representación tradicionales no son capaces de capturar la ponderación, redundancia y escalado simultáneamente. En esta tesis presentamos el marco de representación observacional (Observacional Representation Framework, ORF), el cual supera esta limitación. El ORF integra aspectos de modelos de representación basados en espacios vectoriales, conjuntos de rasgos y teoría de la información. Además, exploramos las implicaciones teóricas y prácticas de ORF de tres formas. En el primer estudio, explotamos ORF como un marco formal para la similitud entre documentos. En este estudio, identificamos las fortalezas y debilidades de las funciones de similitud basadas en espacios métricos (distancia coseno, distancia Euclídea, etc.), conjuntos de rasgos (distancia Jaccard, distancia Dice, etc.) y teoría de la información (pointwise mutual information (PMI), similaridad de Lin, probabilidad condicional, etc.). Para superar las limitaciones observadas en el análisis, definimos el *Information Contrast Model* (ICM), el cual es una generalización parametrizada de PMI. En el segundo estudio, comprobamos empíricamente la habilidad de ORF para integrar características heterogéneas (i.e., carcterísticas con valores discretos y contínuos) sin requerir supervisión. Llevamos a cabo los experimentos en el contexto de clustering de mensajes para la gestión de la reputación online. Finalmente, en el tercer estudio, analizamos ORF como una base formal de fusión de rankings. Nuestro análisis formal muestra que ORF puede ser acomodado a diferentes algoritmos de fusión de rankings dependiendo de las suposiciones hechas, tales como, los algoritmos de esquemas de promedio, Borda, Copeland y Unanimous Improvement Ratio (UIR). Nuestros experimentos en seis conjuntos de datos de fusión de rankings arrojan luz sobre qué aspectos del escenario determinan la idoneidad de diferentes asunciones y algoritmos de fusión de rankings.

Contents

List of Figures

List of Tables

Abbreviations

tf.idf	**t**erm **f**requency - **i**nverse **d**ocument **f**requency
BoW	**B**ag **o**f **W**ords
DUC	**D**ocument **U**nderstanding **C**onferences
ICM	**I**nformation **C**ontrast **M**odel
IC	**I**nformation **C**ontent
IQ	**I**nformation **Q**uantity
IR	**I**nformation **R**etrieval
LDA	**L**inear **D**iscriminant **A**nalysis
LSI	**L**atent **S**emantic **I**ndexing
MRR	**M**ean **R**eciprocal **R**anking
MT	**M**achine **T**ranslation
NIST	**N**ational **I**nstitute of **S**tandards and **T**echnology
NLP	**N**atural **L**anguage **P**rocessing
OIQ	**O**bservational **I**nformation **Q**uantity
ORF	**O**bservational **R**epresentation **F**ramework
ORM	**O**nline **R**eputation **M**onitoring
PCA	**P**rincipal **C**omponent **A**nalysis
PMI	**P**ointwise **M**utual **I**nformation
PoS	**P**art **o**f **S**peech
RepLab	**Rep**utation **Lab**oratory
RTE	**R**ecognizing **T**extual **E**ntailment

SemEval **Sem**antic **Eval**uation

SIM **S**imilarity **I**nformation **M**onotonocity

STS **S**emantic **T**extual **S**imilarity

TE **T**extual **E**ntailment

TREC **T**ext **Re**trieval **C**onference

UIR **U**nanimous **I**mprovement **R**atio

WePS **We**b **P**eople **S**earch

Word2Vect Skip-gram with negative sampling

1 Introduction

1.1. Motivation

Information access is a research area that involves many tasks, such as text mining, information retrieval and text categorization. In all these tasks, document representation is a key step. Some document features are binary, such as word occurrence, named entities, links, or any kind of linguistic structure. Other features are defined in continuous ranges, such as time stamps, topicality, and sentiment polarity.

We highlight three main issues in document representation. First, different features have different levels of importance in the information access process. For instance, the word "Obama" has more weight than "said" when managing news. The second issue is the analysis of feature dependencies. For instance, expected words do not provide new information. In the news domain, "Barack Obama" does not contribute substantially to the information provided by "Obama", given that "Obama" is sufficiently informative in this context. The third issue is feature scaling. For instance, time stamps and word occurrences are expressed on completely different scales.

These three issues are addressed in different ways in supervised or unsupervised scenarios. In the former case, manually annotated output samples are available, and features can be weighted, reduced or projected on the basis of their predictive power. In other words, the training process adapts the learned model to the statistical dependencies and scaling properties of the features. For instance, a supervised classifier will learn that "Obama" is more relevant than "said" when classifying news by topic. It can also infer that "Barak" does not provide additional information relative to "Obama" and that news published less than 72 hours ago is more relevant to readers than older news. Although supervised approaches have been shown to be highly effective in some contexts, their drawbacks have also been widely discussed in the literature, such as overfitting, domain dependency, data bias, and a high annotation cost. Another important drawback is that supervised learning does not provide mechanisms for managing pieces of information, e.g., aggregation or comparison operators.

On the other hand, in the absence of human-annotated data, the weighting, dependence or scaling of features is determined in accordance with their distribution in a document

collection. Typically, unexpected features have greater presence in a representation than expected feature values. For instance, the word feature "Obama" has greater weight in a document representation than frequent common words. Feature dependencies can also be inferred from co-ocurrence. For instance, "Barack" and "Obama" are two word features that tend to appear together. As we will analyse in later chapters, the unexpectedness and co-ocurrence of features serve as the basis of the popular tf·idf feature weighting, stopword removal and word sequence perplexity in language models. However, this paradigm is not compatible with the management of continuous feature values. The reason is that estimating expectedness in terms of occurrence requires some kind of value discretization. That is, although we can estimate the probability of a word, n-gram, tag, etc., the likelihood of a time stamp depends on the granularity with which time is discretized (e.g., days, minutes, etc.). To our knowledge, there are no standard criteria for quantifying the likelihood of continuous feature values in the context of information access.

To overcome this challenge, this thesis presents the Observational Representation Framework (ORF). This approach integrates properties from representation frameworks based on feature sets, vector spaces and information theory. Similar to vector space representations, it captures continuous values. Similar to feature-set-based representations, it allows the application of operators such as inclusion or union, and similar to information-theory-based representations and weighting functions, the ORF weights features in terms of their likelihood.

The ORF has relevant implications from various perspectives. In this thesis, we delve into three of them. First, it provides a common theoretical framework for analysing, comparing and generalizing document similarity functions that are based on different representation schemes. Second, it allows the integration of intrinsic and extrinsic document features in the same representation. Intrinsic features include words, n-grams, etc., whereas extrinsic features may be the output of a clustering process or category membership values generated by classification systems. Third, the ORF provides us with a theoretical foundation and mechanism for ranking fusion.

1.2. Contributions

The first contribution in this thesis is an in-depth study of the benefits and limitations of existing representation models. In particular, we analyse their ability to capture feature specificity, diversity and quantitativity (discrete vs. continuous feature values). After formalizing a number of desirable properties, we observe that none of the families of document representation frameworks (e.g., representations based on sets, metric spaces, language models, etc.) complies with all constraints.

On the basis of this analysis, the second and main contribution of this thesis is the definition of the Observational Representation Framework (ORF), which extends Shannon's traditional notion of the information content ($-log(P(x))$) to the management of continuous feature values. This extension is called the Observational Information Quantity

(OIQ) and is grounded on fuzzy feature sets and inclusion relationships between document observation outcomes in a document collection. We study the formal properties of the ORF and OIQ in a comprehensive way as well as their generalization power relative to traditional representation approaches.

The third contribution is an analysis of similarity functions and their foundations (i.e., cosine distance, Euclidean distance, feature overlap, etc.). We will see, through a study of counterexamples and evidence provided in the literature, that neither Euclidean axioms nor set-based axioms (Tversky's model) properly capture similarity in the context of information access systems. On the basis of the ORF, we review the axiomatic on which traditional similarity functions are based. Again, our analysis shows that different families of similarity functions comply with different constraints. Based on this analysis, we present a general and parametrizable similarity function called the Information Contrast Model (ICM). The ICM, in addition to satisfying desirable formal constraints, generalizes traditional functions such as the PMI, conditional probability, Euclidean distance and Tversky's linear contrast model.

The fourth contribution is related to the capability of the ORF to aggregate heterogenous features into a document representation. For this, we develop a study case: the clustering of tweets in the context of online reputation management. We empirically prove that the model effectively integrates discrete features (words) with continuous feature values. In our study case, the continuous values represent the proximity to pre-annotated categories of tweets and previously generated clusters. The results show that adding heterogeneous features increases the predictive power regarding the similarity between tweet representations. In this sense, the ORF allows us to integrate explicit features (i.e., words) with features extracted from supervised processes (class membership).

Finally, the fifth contribution is a study of the foundations of unsupervised ranking fusion on the basis of the OIQ and ORF. The application of our framework to ranking fusion is developed on the basis that rank scores can be interpreted as quantitative document features. We verify that the Observational Information Quantity (OIQ) generalizes traditional ranking fusion algorithms and explains the effectiveness of existing approaches in different situations. We empirically study these phenomena in six different ranking fusion scenarios.

1.3. Structure of the Thesis

This thesis is organized in eight chapters. A brief summary of the content of each chapter is provided below.

Chapter 1

> **Introduction**: It provides a motivation for the formalization of document representation as a task of information access and establishes the contributions of this thesis.

Chapter 2

A Review of Representation Frameworks: We review the main representation approaches in unsupervised tasks. We highlight their strengths ans weaknesses, analysing their ability to capture: (i) specificity, which establishes that the less common aspects of the information pieces should have greater relevance, since they are the features that distinguish them from the rest of the information pieces, (ii) diversity, which establishes the existence of relationships between the different features of the information pieces; the elimination of redundancies facilitates the study of these relationships and (iii) quantitativity, which establishes the need to capture binary and quantitative characteristics.

Chapter 3

Observational Representation Framework: Our representation framework ORF is presented. Our framework deals to an extension of the traditional Shannon's notion of information content, the one we have called *Observational Information Quantity* (OIQ). This extension is able to manage continuous feature values. ORF not only fulfils the three properties highlighted in the previous chapter (specificity, dependence and quantitivity), but also verifies others, such as monotonicity with respect to values and features, as well as monotonicity with respect to union and the combination of inverse features. It is also able to generalize the most used representation models.

Chapter 4

Revisiting Similarity Axiomatics: We present a revision of the similarity axiomatic between pieces of information, such as distances in a metric space, Tversky's feature-based similarity, etc. Based on the hypothesis that there is a universal set of similarity principles that must be observed with respect to the space of features and the representations of pieces of information, we define a set of restrictions: *identity*, *identity specificity*, *unexpectedness* and *dependency*. These restrictions can be summarized in a single axiom: *similarity information monotonocity* (SIM), which considers *pointwise mutual information* (PMI) and conditional probability as two complementary aspects.

Chapter 5

Analysing Similarity Functions: In this chapter, similarity functions are classified according to their representation paradigm. Based on ORF, we propose a similarity measure called information contrast model (ICM). ICM generalizes both the Pointwise Mutual information and the set-based models considering additions and joints of information quantities. We also present a study case on sentence similarities based on statistics in a popular image description corpus.

Chapter 6

On-line Reputation Management on Twitter: A study case: We focus on the reputation-monitoring scenario, in which social media messages are analysed to identify conversations or events that can affect the reputation of a company or brand. The proposed ORF model is compared with different representation frameworks, using as baseline common schemes, such as bag of words and *tf.idf*. In order to measure the proximity between information pieces, similarity measures common in the literature are used (pointwise mutual information, Jaccard and Lin's distances), in addition to the similarity measure proposed in this work: ICM. Our experiments confirm the hypothesis that adding heterogeneous features under the same ORF-based weighting criterion increases progressively the similarity estimation performance, even when features include both discrete and continuous values and have different scale properties. Finally, a small study is carried out to improve the performance of the approaches through the parameterization of the proposed model.

Chapter 7

Applying the Observational Information Framework to Ranking Fusion: Based on experimental results, we highlight a set of desirable properties that any ranking fusion procedure should satisfy. We then analyse whether the main ranking fusion methods, such as averaging, Borda's rule, the family of Condorcet's methods, etc, satisfy them. Then, we observe that the ORF model presented in this work can be adapted as a ranking fusion method (assuming item scores as features). In addition, ORF satisfies all the desired properties, and moreover, we see under which conditions the ranking fusion algorithms approximate OIQ. Finally, we also present the performance of the ranking fusion methods in the experimental part.

Chapter 8

Conclusions: A summary of each chapter can be seen, and some conclusions are drawn.

In addition, the thesis contains an appendix with the formal demonstrations of the statements established in previous chapters.

Part I

Representation Frameworks

2

A Review of Representation Frameworks

2.1. Introduction

In this chapter, we analyse existing document representation frameworks. Representing documents in non-supervised scenarios is challenging. In the absence of training data sets, it is not easy to weight or scale feature values. For this purpose, a representation must capture at least three aspects of feature. The first is *specificity*. Uncommon features, such as infrequent words, have greater weight than common features. This principle underlies many popular representation techniques, such as *tf·idf* weighting, stopword removal and perplexity in language models. The second aspect is *feature dependence*. Statistically redundant features do not add information about documents. This principle underlies many dimensionality-reduction methods such as latent semantic indexing, latent Dirichlet allocation and word embeddings. The third property is *quantitativity*. A suitable representation framework should be able to capture both discrete and continuous feature values (e.g., binary word occurrence vs. class membership or latent features). These three aspects will guide our analysis throughout this chapter.

Throughout this analysis of the literature, we will see that Shannon's notion of information content (IC), which is expressed as $I(x) = -\log(P(x))$, is an underlying concept in many representation schemes, enabling the capture of *specificity* and (in some cases) *dependency*. However, the traditional definition of the IC is not compatible with continuous feature values (*quantitativity*), given that the unlikelihood of features is estimated in terms of discrete statistical events. Consequently, existing representation models are not able to simultaneously capture *specificity*, *dependency* and *quantitativity*.

2.2. Documents as Vectors: Vector Space Models

In the following, we consider a collection of documents, denoted by $\mathcal{D}$, and a set of features, denoted by $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$. We assume that for each document, there exists a function, namely, $\pi_d(\gamma_i)$, that projects the document onto each feature γ_i. We formalize existing document representation models on the basis of this function.

Table 2.1: Characterization of representation models. None of the existing representation models are able to capture the three properties simultaneously.

	Specificity	**Dependency**	**Quantitativity**
Documents as a Vector Space			
tf (1)	–	–	–
tf.idf (2)	✓	–	–
Documents as Density Distr.			
Normalized tf (3)	–	–	✓
Documents as Feature Sets			
Boolean model (4)	–	–	–
Multi-sets (5)	–	–	–
Lin's Model (6)	✓	–	–
Documents as Fuzzy Sets			
Original Fuzzy Set (7)	–	–	✓
Luca and Termini, Kaufman (8)	–	–	✓
Zadeh's model (9)	✓	–	–
Documents as Feature Seq.			
Non probabilistic (10)	–	–	–
General Language Model (11)	✓	✓	–
n-grams Language Model (12)	✓	–	–
Neural Language Models (13)	✓	✓	–
Documents as Single Events:			
Conjoint of Features (14)	✓	✓	–
Documents as Feature Vectors with Dimensionality Reduction			
Principal Component Analysis (15)	–	✓	✓
LDA, LSI (16)	–	✓	–

In vector space representation models, each document, $d \in \mathcal{D}$ is projected into an n-dimensional space in accordance with its feature values:

$$\boxed{d = (x_1, \ldots, x_n), \qquad x_i = \pi_d(\gamma_i), 1 \leq i \leq n}$$

The simplest representation approach is the binary bag of words (BoW) representation, in which the value of the projection function, namely, $\mu_d(\gamma_i)$, is either zero or one depending on whether the target feature is present in the document. This model can be extended to the term frequency (*tf*), which considers the frequency of words or terms instead of their appearance; in this case, $\pi_d(\gamma_i) \in \mathbb{N}$. To capture the *specificity* of features, vector space models are complemented by a weighting factor; the most popular model of this kind is *tf·idf* (see (2) in Table 2.1):

$$\pi_d(\gamma_i) = tf(d, \gamma_i) \cdot idf(\gamma_i) \ ,$$

where:

$$tf(d, \gamma_i) = \text{Frequency of } \gamma_i \text{ in d} \ ,$$

$$idf(\gamma_i) = log\left(\frac{|\mathcal{D}|}{|\{d' \in \mathcal{D} : tf(d', \gamma_i) > 0\}|}\right)$$

There is a direct correspondence between *idf* and the concept of IC ($IC(\gamma_i) = -log(P(\gamma_i))$ [101]. Thus, vector space models incorporate the notion of the IC to weight features in accordance with their *specificity*.

Several extensions of idf weighting have been proposed, such as x^I [100], the z-measure [54], the residual idf [23] and gain [93]. More recently, Shirakawa et al. [108] proposed an extension of idf for longer text pieces, such as phrases, that is based on information theory. The use of Shannon's entropy in IR has been explored in terms of weighting, feature selection and search engines [64, 53, 61]. In all these cases, the feature weighting methods focus on discrete features (e.g., words or phrases).

Another extension of vector space models consists of representing documents as probabilistic density functions. This establishes the following constraint:

$$\sum_{i=1}^{n} \pi_d(\gamma_i) = 1$$

Within this family, the most common approach is the *normalized tf* approach (see (3) in Table 2.1), which considers the frequencies of words (features) relative to the total number of words in a document. This extension enables the consideration of quantitative features; however, *dependency* and *specificity* are not captured since the distribution of features in the whole collection of documents is ignored.

A general limitation of vector space models, e.g., (1), (2) and (3) in Table 2.1, is that they do not consider feature *dependency*. For instance, single words in common phrases (e.g., "Obama" vs. "Barack Obama") can be redundant. There are two main types of approaches for mitigating this issue. The first consists of considering more complex linguistic units, such as named entities, n-grams, and phrases, as features. The second consists of applying dimensionality-reduction processes, which are discussed in later sections. Another important drawback, which is the focus of this thesis, is that *specificity* (i.e., *tf·idf*) is formalized on the basis of the feature frequency in documents; hence, it is not possible to manage continuous valued features (*quantitativity*).

2.3. Documents as Feature Sets

Documents can be represented as sets of features. The most basic approach is the Boolean model, shown with (4) in Table 2.1, where $\gamma_i \in \Gamma$ is a feature in consideration of the following:

$$\boxed{d = \{\gamma_1, \ldots, \gamma_m\}, \quad \gamma_i \in \Gamma, \quad 1 \leq i \leq m \leq n}$$

Words and n-grams are the most commonly considered features; however, more complex linguistic structures or even semantic units such as ontological concepts can also be used in such a representation. Feature-set-based frameworks enable the application of set operators, such as the intersection and union operators. Accordingly, it is easy to adop set-theoretic similarity measures, such as the Jaccard distance: $\left(\frac{|d \cap d'|}{|d \cup d'|}\right)$. These frameworks can be extended to countable features by using multi-sets and applying the

maximum and minimum operators as the union and intersection operators; see (5) in Table 2.1.

The main disadvantage of feature-set-based representation frameworks is that they do not address *specificity* and *dependency*. In general, weights of the features correspond to their salience in a document, whereas these two properties require consideration of the statistics of the features in the whole document collection. However, *specificity* can be captured by appending probabilities to a feature-set-based representation model; see (6) in Table 2.1. That is, it is assumed that the features are statistically independent. Therefore, the specificity and IC of a document can be estimated as products of probabilities across features. In addition, it is possible to compute the union and intersection of documents by applying the same operators as in set-based models:

$$d = \{\gamma_1, \ldots, \gamma_m\}, \quad P\big(\{\gamma_1, \ldots, \gamma_m\}\big) = \prod_{i=1}^{m} P(\gamma_i) \ .$$

$$P(d \cap d') = \prod_{\gamma \in d \cap d'} P(\gamma), \qquad P(d \cup d') = \prod_{\gamma \in d \cup d'} P(\gamma) \ .$$

Dekang Lin [77] used such a representation to estimate the similarity between pieces of information in terms of their unique and common contents. However, this representation sacrifices *dependency*. Moreover, as in weighted vector space models, the definition of the IC in Shannon's theory is not able to capture quantitative features.

2.4. Documents as Fuzzy Sets

A natural way to extend feature-set-based models to quantitative features is via *fuzzy sets*; see (7) in Table 2.1. Consider a collection of features Γ, which is also called a *universe* of features. A document d can be directly represented as a fuzzy set as follows:

$$\boxed{d = \big(\Gamma, f\big), \quad f(\gamma_i) = \pi_d(\gamma_i)}$$

where $\mu_d(\gamma_i)$ represents the salience or membership degree of feature γ_i with respect to document d. Corresponding to the ordinary set operations of union and intersection, fuzzy sets have similar operations that are based on maxima and minima. Fuzzy sets were applied as an extension of the basic Boolean model in the 1980s and 1990s [26]. However, although fuzzy-set-based representations can capture quantitative features, *specificity* and *dependency* are still ignored. Similar to the weighting schemes in vector space models, capturing *specificity* requires linking fuzzy-set-based representations to information theory.

Decades ago, various proposals asserted a direct connection between fuzzy sets and Shannon's entropy. For instance, [33] and [63] transformed a fuzzy representation into a probability density function; see (8) in Table 2.1. Then, they computed the entropy of the corresponding document in the traditional manner. Although this representation captures *quantitativity*, like representation models based on density distributions (see

the previous section), it is not able to capture *specificity* and *redundancy* given that the distribution of features across documents is not inferred; only the distribution, within the same document is considered.

Another exception is a model that was previously proposed by [122]; see (9) in Table 2.1. The author suggested a definition that considers both the probabilities and memberships of elements.

$$H(d) = -\sum_{\gamma_i \in \Gamma} \pi_d(\gamma_i) \cdot P(\gamma_i) \cdot log\big(P(\gamma_i)\big) \ .$$

This representation approach captures *specificity* by means of the $P(\gamma_i)$ component, which represents the likelihood of a feature in probabilistic terms. However, this comes at the cost of discretizing the features (*quantitativity*). In addition, this model assumes feature independence.

A connection between fuzzy set-based representations and the IC has been partially established, in a theory referred to as *fuzzy information theory* [67, 104]; however, this theory focuses on the *vagueness* of fuzzy sets rather than on their unlikelihood (specificity) in the space of pieces of information.

Other approaches extract *linguistic variables* from documents and represent them as fuzzy sets [35, 55]. However, these approaches are outside the scope of this thesis, as they do not focus on document representation.

2.5. Documents as Sequences of Features: Language Models

Other representation models start from the assumption that documents are sequences of features (words or characters); see (10) in Table 2.1. Let γ_d^i denote the feature that is located in the i-th position of document d, and let m denote the document length:

$$\boxed{d = \big(\gamma_d^1, \ldots, \gamma_d^m\big), \qquad \gamma_d^i \in \Gamma, \ \ 1 \leq i \leq m}$$

The number of dimensions corresponds to the length of the document rather than the number of features (as in a vector space model). Most editing-distance-based measures work under such a representation [89, 112]. They are based on the number of changes that are necessary to transform one sequence of words or characters into another sequence.

To capture *specificity*, language models extend this representation by considering the probability distribution of word sequences; see (11) in Table 2.1. The inclusion relationships between shorter and longer sequences permit the inference of probabilities of documents of various lengths. A key concept in language models is *perplexity*, which refers to the likelihood of a word sequence normalized by its length. There is a direct correspondence between the IC and perplexity. Texts are word sequences, and the

perplexity of a text in the context of a language model reflects its IC.

$$Perplexity(d) = P\big(\gamma_d^1, \ldots, \gamma_d^m\big)^{-\frac{1}{m}} = \prod_{i=1}^{m} P\big(\gamma_d^i \mid \gamma_d^1, \ldots, \gamma_d^{i-1}\big)^{-\frac{1}{m}} \ .$$

$$\mathcal{I}\big(\gamma_d^1, \ldots, \gamma_d^m\big) = -log\Big(P\big(\gamma_d^1, \ldots, \gamma_d^m\big)\Big) = m \cdot log\Big(Perplexity\big(\gamma_d^1, \ldots, \gamma_d^m\big)\Big) \ .$$

The main contribution of language models is that they capture both word specificity and dependencies. Language modelling has been exploited in multiple tasks. Several works have expanded such representations to other linguistic features, such as part of speech (PoS) tags and dependency structures [113]. Zhai [123] showed that a connection exists between language models and the *tf·idf* representation technique. However, since language models are based on word occurrence, they do not capture quantitative features.

On the other hand, estimating the likelihoods of long sequences of words is highly challenging. Typically, this challenge is addressed via the n-gram model; see (12) in Table 2.1, which assumes independence beyond a certain distance n: $P\big(\gamma_d^1, \ldots, \gamma_d^m\big) \simeq \prod_{i=1}^{n} P\big(\gamma_d^i \mid \gamma_d^{i-n}, \ldots, \gamma_d^{i-1}\big)$. This approach partially captures *dependency* (only for contiguous words).

A fundamental problem in this paradigm is the curse of dimensionality, which limits modelling on larger corpora for universal language models. *Neural language models* are language models based on neural networks, exploiting their ability to learn distributed representations to reduce the impact of the curse of dimensionality. The first approaches developed in this direction were word embeddings, which operate under the assumption that the meaning of a word is defined by its textual context. In particular, neural embeddings represent words as their internal neural network representations [85, 95]. Some previous works [11, 74] have demonstrated that the most popular word embedding model, namely, skip-gram with negative sampling (Word2Vec) implicitly factorizes a word-context PMI matrix: $\langle \vec{v}_w, \vec{v}_{w'} \rangle \approx PMI(w, w')$. The main consequence is that a correspondence exists between the inner scalar product of such word representations and the IC: $\langle \vec{v}_w, \vec{v}_w \rangle \approx \mathcal{I}(w)$. That is, the word vector length in Word2Vec preserves the IC of words, thereby capturing *specificity*. Linguistic units that are longer than words and other features, such as PoS tags and topic identifiers, have also been used [107].

Word embeddings generate word representations. However, in recent years, multiple contextual neural language models have emerged that provide representations, not only for single words but also for word sequences. These leverage the intuition that the meaning of a particular text depends not only on the identity of a word itself, but also on the other words that surround it this moment. Some models that have had particular impact in the community are LSTM [58], COVE [81], ELMo [96], ULMFit [59] and BERT [36].

2.6. Documents as Single Statistical Events

Another alternative approach consists of representing documents as sets of features with a conjoint probability distribution over the whole collection of documents:

$$\boxed{d = \{\gamma_1, \ldots, \gamma_n\}, \quad P(d) = P(\gamma_1, \ldots, \gamma_n)}$$

This representation approach is able to capture the *specificity* of features, given that less frequent features affect the conjoint probability to a greater extent. That is, the conjoint probability is upper bounded by the probability of single features:

$$P(\gamma_1, \ldots, \gamma_n) \leq P(\gamma_i), \quad 1 \leq i \leq n \ .$$

In addition, unlike previous approaches, this representation model captures *dependency*. For instance, redundant features do not affect the likelihood of documents:

$$P(\gamma_1, \ldots, \gamma_i, \gamma_i, \ldots, \gamma_n) = P(\gamma_1, \ldots, \gamma_i, \ldots, \gamma_n), \quad 1 \leq i \leq n \ .$$

This representation enables the application of information-theory-based similarity measures. In particular, the pointwise mutual information (PMI) of two documents, denoted by d and d', has been widely used. The PMI can also be expressed as a combination of ICs:

$$PMI\big(d, d'\big) = log\left(\frac{P\big(d, d'\big)}{P(d) \cdot P(d')}\right) = \mathcal{I}\big(d\big) + \mathcal{I}\big(d'\big) - \mathcal{I}\big(d, d'\big) \ ,$$

where $\mathcal{I}(d) = -log\big(P(d)\big)$ and $\mathcal{I}\big(d, d'\big) = -log\big(P(d, d')\big)$. Moreover, the IC can be expressed as the PMI of a document with itself: $PMI(d, d) = \mathcal{I}(d)$. The PMI has been proven to be highly effective in multiple word-similarity tasks [20, 28]. In addition, several works on state-of-the-art approaches combine the notion of the IC with the topological depth of ontologies [98, 1].

A conjoint probabilistic representation, which corresponds to (14) in Table 2.1, has the main limitation of capturing only discrete feature values. Again, the reason is that the notion of the IC was originally defined over discrete statistical events, thereby sacrificing *quantitativity*.

2.7. Dimensionality Reduction

According to our analysis, the existing models do not capture *dependency* or provide a feasible way of estimating it. According to the literature, the most successful way of overcoming this problem is by projecting an object representation into a reduced set of dimensions:

$$\boxed{d = f_{proj}\big(\pi_d(\gamma_1), \ldots, \pi_d(\gamma_n)\big), \quad f_{proj} : \mathbb{R}^n \longrightarrow \mathbb{R}^m, m \ll n}$$

where f_{proj} is a projection function that reduces the dimensionality from n to m. The available projection functions include principal component analysis (PCA), kernel PCA, and linear discriminant analysis (LDAn). PCA avoids redundancy and allows the management of continuous feature values; see (15) in Table 2.1. However, the notion of specificity for continuous feature values remains an open issue.

In the context of text representation, the most traditional and well-known approach is latent semantic indexing (LSI). As successors to LSI, in the 2000s, generative topic models such as latent Dirichlet allocation (LDA) [18] became popular; see (16) in Table 2.1. LDA-based models have been exploited in multiple unsupervised scenarios. Given the generalization power of generative models, some attempts have been made to include continuous feature values in the Topic Over Time model [117]; however, the complexity of the model definition does not permit the incorporation of many quantitative features. Note that the neural language models described in Section 2.5 can also be understood as dimensionality reduction techniques.

2.8. Conclusions: The Gaps in Representation Models

The analysis described in this chapter suggests that the IC is a key issue in many representation models. It has direct correspondences with the popular *idf* weighting method, the *perplexity* in language models, information-based measures such as the PMI and some neural language models. The notion of the IC captures both *specificity* and (at least on a theoretical level) *dependence*. However, Shannon's IC does not enable the management of continuous feature values. The reason is that it requires the estimation of probabilities of discrete events. Note that the traditional *differential entropy* quantifies the information of a whole distribution. However, the information quantity of a single event (IC) cannot be applied to single values in a continuous distribution.

Consequently, *specificity*, *dependence* and *quantitativity* are incompatible in existing representation models (see Table 2.1). Therefore, existing representation models do not provide mechanisms for integrating, without supervision, heterogeneous features such as time or second level features generated by processing tools (polarity, topicality, factuality, centrality, etc.). In this sense, the main goal of the representation model proposed in this thesis (the Observational Representation Framework) is to allow discrete and continuous-valued features to be combined in the same representation. The proposed representation approach is described in the next chapter.

3

Observational Representation Framework

3.1. Introduction

Measuring an object of study is a fundamental aspect of research in many disciplines. Information systems and, in particular, information retrieval and natural language processing, are not exceptions. In these areas, information is the object of study. The closest notion to a measure of information is Shannon's information content (IC), which quantifies the information of a single message m in terms of its probability of being drawn from among all possible choices in the message space: $I(m) = -log(P(m))$. As shown by our analysis in the previous chapter, the IC is a core concept in many representation frameworks. Features that add specificity to documents have more weight than frequent or statistically redundant features. We referred to those aspects of features as *specificity* and *dependence*.

However, Shannon's notion of the IC has an important limitation. As seen in the previous chapter, the IC is applicable only to discrete events. Therefore, messages cannot be characterized by continuous features[1]. That is, the IC is not compatible with quantitative features such as the document length, creation date or number of views. Other quantitative features of documents include the outputs of external system components, such as the response of a sentiment polarity classification module or the expected relevance according to alternative search engines. Features of this kind cannot be captured and measured with the traditional notion of the IC. In the previous chapter, we referred to this capability of a representation framework as *quantitativity*.

In this chapter, we present the *Observational Representation Framework*, which is grounded on the *Observational Information Quantity* (OIQ) and its properties. The OIQ generalizes the definition of Shannon's IC to continuous values. The OIQ quantifies information as in language models, captures quantitative features as in vector space models, and allows an information algebra to be defined via set operators as in set-based representations.

In Section 3.2, the proposed observational representation model is defined and the al-

[1] Notice that *differential entropy* quantifies the information of continuous distributions, but not single values.

lowed algebraic operations are specified; in Section 3.3, the formal properties of this model are analysed. Section 3.4 addresses the computability of the OIQ under various statistical assumptions. Finally, in Section 3.5, the OIQ is related to several traditional methods of representing texts: the weighted vector space model, language models and distributional representations among others.

3.2. Defining the Observational Information Quantity

The *Observational Information Quantity* (OIQ) [8, 48] is based on the intuition that computers observe feature sets rather than documents themselves. Consequently, the OIQ framework starts from the assumption that each document is unique and that there exists a practically infinite amount of information underlying each document. This implicit information is determined by aspects such as the author, context, situation, time, scenario, channel, etc. In other words, all documents are extremely and equally improbable. Therefore, we cannot compare them to each other in terms of the traditional IC ($-log(P(m))$).

The observational framework studied in this thesis solves this conflict through a change of paradigm. In brief, documents have no measurable information quantity to be estimated, but the feature set instantiation can be derived from the observed document outcome. Traditional Shannon information theory quantifies the IC in terms of the probability of the message of interest being drawn from among all possible choices in the message space. In contrast, the **OIQ quantifies the IC of a message in terms of the probability of its features appearing (being subsumed) in other messages. Consequently, the OIQ quantifies to what extent the observed outcome allows the message to be identified within the space of all messages.**

To capture quantitative features, document observation outcomes are modelled as fuzzy sets, and their inclusion relationships are used to model their likelihood, that is, the number of document observations in which a certain observation is subsumed. In this way, using inclusion relationships between representations based on fuzzy feature sets, we can capture quantitative features and measure their informativeness in a probabilistic framework.

Let $\mathcal{D}$ be the countable and infinite set of possible documents and let Γ be a set of features. The projection of each document, $d \in \mathcal{D}$, onto a feature, $\gamma \in \Gamma$, is denoted, as in Chapter 2, by $\pi_d(\gamma)$. That is, as in vector space models, the representation is grounded on the feature *salience* in documents. Then, document observation outcomes are modelled as *fuzzy feature sets*, which are instantiations of quantitative features.

Definition 3.1 (Observation Outcome). *An observation outcome X under the universe of features Γ is a fuzzy set (Γ, f) of features, where $f : \Gamma \to \mathbb{R}^+ \cup \{0\}$ is a feature membership function.*

This means that an observation outcome is basically a value instantiation (membership function) of features. On the other hand, each document has an associated Document Observation Outcome:

Definition 3.2 (Document Observation Outcome). *The observation outcome of a document d, denoted by $\mathcal{O}_\Gamma(d)$, under the universe of features Γ is an observation outcome $\big(\Gamma, \pi_d\big)$, where the membership function, $\pi_d : \Gamma \to \mathbb{R}^+ \cup \{0\}$ corresponds to the document feature projection function.*

The purpose of using the notion of fuzzy sets is to exploit the associated inclusion operators (see Definition 3.3). Regarding the range of a membership function, it should be noted that the image of a feature set can be normalized; thus, the range of salience or membership values is not relevant to our study.

According to the above definitions, the notion of an *observation outcome* can be understood independently of the notion of a document. In fact, we can say that any fuzzy set of features, denoted by X, is a possible observation outcome. This understanding allows us to treat an observation outcome as an abstract object, without thinking in terms of the representation of a particular document. For clarity, we will denote the fuzzy feature set $X = \big(\Gamma, f\big)$ by $(x_1, \ldots, x_n)$ where $x_i = f(\gamma_i)$ and $n = |\Gamma|$.

Now, we formalize the occurrences of an observation outcome X as follows.

Definition 3.3 (Occurrence of an Observation Outcome). *The occurrence, $Occ(X)$, of an observation outcome X is the set of documents within the document set $\mathcal{D}$ whose observation outcomes subsume X.*

$$Occ(X) = \big\{d \in \mathcal{D} : \mathcal{O}_\Gamma(d) \supseteq X\big\} \ .$$

To capture information-theory-based similarity functions such as the pointwise mutual information or Lin's similarity, it is necessary to define a formal probabilistic framework. Our framework considers the probabilistic space $\big(\mathcal{D},\ \wp(\mathcal{D}),\ P\big)$, where the probability function is defined over the power set of documents $\wp(\mathcal{D})$.

Under this probabilistic framework, an occurrence set $Occ(X)$ is actually an element of the power set $\wp(\mathcal{D})$, i.e., a probabilistic event:

$$P\big(Occ(X)\big) = P\big(\{d \in \mathcal{D} : \mathcal{O}_\Gamma(d) \supseteq X\}\big) \ .$$

Therefore, we can model the likelihood of a fuzzy feature set as its probability of occurrence among document observation outcomes. For instance, if words are considered as features (BoW), the text (or document) $d =$ *"My pen"* produces the observation outcome $\mathcal{O}_\Gamma(d) = \big(\{$*"My"*, *"pen"*$\}, f\big)$. The occurrence of $\mathcal{O}_\Gamma(d)$ is the set of documents containing *"My"* and *"pen"*, which is a superset of the occurrence of *"My little pen is lost"*:

$$Occ\big(\{\text{“My”}, \text{“pen”}\}\big) \supset Occ\big(\{\text{“My”}, \text{“pen”}, \text{“is”}, \text{“lost”}\}\big) \ .$$

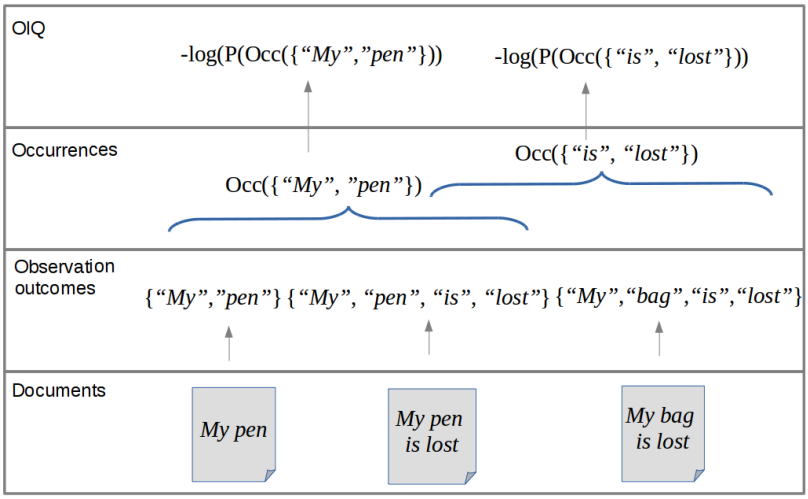

Figure 3.1: An example of documents, observation outcomes, occurrences and observational information quantity over the BoW feature set.

Thus, *larger* observation outcomes have lower likelihoods than *smaller* outcomes.

$$P\Big(Occ\big(\{\text{“}My\text{”}, \text{“}pen\text{”}\}\big)\Big) > P\Big(Occ\big(\{\text{“}My\text{”}, \text{“}pen\text{”}, \text{“}is\text{”}, \text{“}lost\text{”}\}\big)\Big) \ .$$

Note that our sample space is defined by the set of possible documents $\mathcal{D}$, while traditional probabilistic representation frameworks are defined over a feature sample space $\big(\Gamma, \wp(\Gamma), P\big)$, which prevents the modelling of quantitative features (probabilistic events are discrete). In addition, considering documents as the sample space allows us to capture the notion of feature occurrence as well as statistical relationships between features via set operators.

The ORF combines strengths from various representation frameworks. Similar to feature-set-based representations, it allows the union and intersection operators to be applied to observation outcomes. Similar to vector-space-based representations, it allows continuous feature values to be captured via fuzzy sets. Similar to probabilistic representation frameworks, it allows the likelihood of features to be defined. This is the purpose of the OIQ which is defined as follows:

Definition 3.4 (Observational Information Quantity). *The Observational Information Quantity (OIQ) of an observation outcome, denoted by $\mathcal{I}\big(X\big)$, is the negative logarithm of the probability of its occurrence:*

$$\mathcal{I}\big(X\big) = -log\Big(P\Big(Occ(X)\Big)\Big).$$

For clarity, we will use $\mathcal{I}_\Gamma\big(d\big)$ to denote the OIQ of a document observation outcome under the feature set Γ: $\mathcal{I}_\Gamma\big(d\big) = \mathcal{I}\big(\mathcal{O}_\Gamma(d)\big)$.

We can illustrate the representation framework with the example showed in Figure 3.1. Let us imagine a universe of only three documents *“My pen”*, *“My little pen is lost”* and *“My bag is lost”*. Then, the OIQ that is associated with *“My pen”* would be estimated as

$-log\left(\frac{2}{3}\right)$, given that these two word features are contained in two existing observation outcomes. In a document corpus, the OIQ of an observation of two specified words corresponds to the probability of observing those two words in a randomly selected message.

Note that this formalization links the OIQ with the extent to which the representation identifies the message. In other words, the more information the observation outcome contains, the more likely this message is to be the one that is actually observed. For instance, the words "my" and "pen" appear in multiple of the above messages. If we include additional information such as other words, time, author, or context, then the resulting feature set will identify the particular message.

In terms of the set operators, an inclusion relationship between representations is related to the minimum feature value between them. Therefore, the inclusion of a fuzzy feature set X in a document observation outcome, $\mathcal{O}_\Gamma(d)$ can be expressed as:

$$\mathcal{O}_\Gamma(d) \supseteq X \quad \Leftrightarrow \quad \pi_d(\gamma_i) \geq x_i, \quad i = 1, \ldots, |\Gamma| \ .$$

In other words, the occurrence of an observation outcome X is the set of documents that have a greater instantiation of features $\pi_d(\gamma_i)$ than x_i for every feature. On this basis, we can express the OIQ of an observation outcome X as follows:

$$\mathcal{I}\big(X\big) = log\left(\frac{1}{P\big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq x_i, \quad i = 1, \ldots, |\Gamma|\}\big)}\right) \ .$$

where $\pi_d(\gamma)$ is the projection of features onto documents in $\mathcal{D}$. Likewise, the OIQ of a document observation outcome can be expressed as:

$$\mathcal{I}_\Gamma\big(d\big) = log\left(\frac{1}{P\big(\{d' \in \mathcal{D} : \pi_{d'}(\gamma) \geq \pi_d(\gamma), \forall\ \gamma \in \Gamma\}\big)}\right) \ .$$

In addition, we can aggregate information via the fuzzy set union operator.

Proposition 3.1. *The OIQ of a union of observation outcomes is equivalent to the OIQ of their maximum feature values. Formally, given two observation outcomes X and Y:*

$$\mathcal{I}\big(X \cup Y\big) = -log\Big(P\Big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq \ \max(x_i, y_i),\ i = 1, \ldots, |\Gamma|\}\Big)\Big) \ .$$

3.3. Formal Properties

For simplicity, we illustrate the formal properties of the OIQ by assuming that the documents of interest are texts and the features are words. The first property is monotonicity. The value of the information quantity should grow with the feature values. In the context of word features, increasing the number of occurrences of words in a text should increase its OIQ. This property can be formalized as follows.

Property 3.1 (Feature Value Monotonicity). *Increasing the feature instantiation values increases the OIQ. Given two observation outcomes (fuzzy feature sets) X and Y over a feature set $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$, the following is verified:*

$$x_i \geq y_i, \quad \forall\, i \in \{1, \ldots, n\} \Rightarrow \mathcal{I}\big(X\big) \geq \mathcal{I}\big(Y\big) \ .$$

Note that this property also holds for any quantitative feature. In addition, since the set of possible documents, $\mathcal{D}$, is countable and infinite, a strict increase in one feature is sufficient to produce a strict increase in the OIQ.

$$x_i \geq y_i, \quad \forall\, i \in \{1, \ldots, n\} \wedge \exists i \in \{1, \ldots, n\}\big(x_i > y_i\big) \Rightarrow \mathcal{I}\big(X\big) > \mathcal{I}\big(Y\big) \ .$$

Therefore, the OIQ exhibits *quantitativity*, as it is sensitive to changes in continuous-valued features.

On the other hand, the more features we consider, the more information we have about a document. This aspect is also captured by the OIQ, in the sense that it is also monotonic with respect to the feature set.

Property 3.2 (Feature Set Monotonicity). *Adding features to the set Γ increases the OIQ values of document observation outcomes. Let X and X_{sub} be two observation outcomes such that $X = \big(\Gamma, f\big)$ and $X_{sub} = \big(\Gamma - \{\gamma\}, f\big)$:*

$$\mathcal{I}\big(X\big) \geq \mathcal{I}\big(X_{sub}\big) \ .$$

In addition, the OIQ is additive in regard to the observation outcomes. More concretely, the OIQ is monotonic with respect to the union of observations.

Property 3.3 (Monotonocity of the Union of Observation Outcomes). *Given two observation outcomes X and Y the OIQ of their union is larger than those of the individual outcomes:*

$$\mathcal{I}\big(X \cup Y\big) \geq \mathcal{I}\big(X\big) \ .$$

The above property can be interpreted as follows. Taking union of an increasing number of observation outcomes gradually specifies the identity of the observed document, through either the inclusion of new features or higher feature values. For instance, the union of the two outcomes *{"my","red","pen"}* and *{"red","red","is"}* characterizes the document *"My red pen is indeed red"* to a greater extent than the individual observation outcomes do.

The next property states that the OIQ captures *specificity* by giving more weight to infrequent features. For instance, if two words (features) appear in a document to the same extent, then, as in traditional tf·idf weighting, the information quantity associated with each feature depends on the frequency of that word in the entire collection.

Property 3.4 (Specificity). *The more unexpected a feature instantiation is, the more informative it is. Given two single feature observation outcomes $X = \big(\{\gamma\}, f\big)$ and*

$X' = \big(\{\gamma'\}, f'\big)$ *where* $f(\gamma) = f'(\gamma') = v$*, we have the following implication:*

$$P\big(\{d \in \mathcal{D} : \pi_d(\gamma) \geq v\}\big) < P\big(\{d \in \mathcal{D} : \pi_d(\gamma') \geq v\}\big) \Longrightarrow \mathcal{I}\big(X\big) \geq \mathcal{I}\big(X'\big) \ .$$

The next property is related to *dependency*. Let us consider a collection of web pages. Suppose that *"Barack"* and *"Obama"* always appear together. Then, the probabilities of observing *"Barack"*, *"Obama"* and *"Barack Obama"* on various web pages would be similar. That is, *"Obama"* does not contribute additional information relative to *"Barack"*. This also holds for continuous values. Consider the recentness of web pages as measured in either seconds or days, both of which are real-valued quantities. These two features are redundant in terms of the OIQ. The probability of a page being more recent than another page is independent of whether the recentness is measured in seconds or days. That is, two features are redundant if they are monotonically related.

Property 3.5 (Dependence). *Redundant features do not affect the OIQ values of document observation outcomes. Given two features* γ_1*,* $\gamma_2 \in \Gamma$*, if there exists a real strict monotonic function* g *that satisfies:* $\pi_d(\gamma_1) = g\big(\pi_d(\gamma_2)\big)$*,* $\forall\ d \in \mathcal{D}$*, then,* $\mathcal{I}_{\{\gamma_1\}}\big(d\big) = \mathcal{I}_{\{\gamma_1,\gamma_2\}}\big(d\big)$*,* $\forall\ d \in \mathcal{D}$ *.*

This property is closely related to the *idempotency* axiom in information algebra [65], which states that combining a piece of information with part of itself yields nothing new.

As explained previously, the Observational Representation Framework allows different OIQs to be derived for the same document depending on which features are considered. The OIQ converges to Shannon's IC when an infinite set of features is considered.

Property 3.6 (Convergence with Shannon's Information Quantity). *The OIQ of a document observation outcome under an infinite number of heterogeneous features corresponds to the likelihood of the document itself.*

$$\lim_{|\Gamma| \to \infty} \mathcal{I}_\Gamma\big(d\big) = -\log\big(P(d)\big) \ .$$

This is true whenever we assume that we can always find a feature that discriminates two distinct documents. Therefore, only a document itself improves its feature instantiation in terms of any potential feature. This property suggests that the ground-truth OIQ of a document is obtained as the result of considering an infinite number of features. Therefore, features and their instantiations can be interpreted as artefact to make estimation feasible.

Another possible concern is the effect of combining inverse features. For instance, we could consider both the *'height'* and *'shortness'* of a person. According to the concept of the OIQ, being extremely tall is equally informative as being extremely short. If we combine these two features, the only way that both features can be (non-strictly) satisfied simultaneously is if all people being compared have the same height. For instance, if we consider the occurrence of a word as a feature, the OIQ will correspond to the likelihood of observing this word. By contrast, if we consider the absence of the same word as a feature, then the OIQ will correspond to the likelihood of representations that do

not contain this word. If both the occurrence and absence of a word are considered as features, then the OIQ can be interpreted as the likelihood of observing exactly the same numbers of word occurrences. Formally, this property is described as follows.

Property 3.7. *The OIQ for two inverse and continuous-valued features corresponds to the probability of equality in this feature. Given a document $d \in \mathcal{D}$ and a feature $\gamma \in \Gamma$, consider the set of features $\{\gamma, \gamma^{-1}\}$, where γ^{-1} is defined by $\pi_d(\gamma^{-1}) = \pi_d(\gamma)^{-1}$, $\forall d \in \mathcal{D}$; then:*

$$\mathcal{I}_{\{\gamma,\gamma^{-1}\}}(d) = -\log\Big(P\big(\{d' : \pi_{d'}(\gamma) = \pi_d(\gamma)\}\big)\Big) \ .$$

This property has also been studied in algebraic information theory [65] (the nullity axiom), in which it has been concluded that contradictory information absorbs everything and can only be derived from contradictions.

3.4. Estimating the Observational Information Quantity

According to Definition 3.4, computing the OIQ requires estimating the following:

$$\mathcal{I}_{\Gamma}(d) = log\left(\frac{1}{P\Big(\big\{d' : \pi_{d'}(\gamma) \geq \pi_d(\gamma), \forall\ \gamma \in \Gamma\big\}\Big)}\right) \ .$$

When managing a limited set of document features (e.g., creation date, views, topicality and sentiment polarity), it is possible to estimate the conjoint probability distribution of features from a finite collection $\mathcal{D}^c \subset \mathcal{D}$. We can use conjoint cumulative distributions and copula analysis techniques to overcome this challenge.

On the other hand, in the case of discrete features such as words, OIQ estimation converges with the traditional challenge encountered in both n-grams and neural-based language models, that is, the estimation of the probability of a sequence of discrete features. The next section formally describes this connection.

The open issue is how to integrate categorical and quantitative features. Unfortunately, an exponential number of samples would be required to estimate the conjoint cumulative distribution for all features simultaneously[2]. One way to address this problem is by assuming independence between categorical and quantitative features:

Definition 3.5. Feature Independence: *Given two disjoint sets of features, $\mathcal{F}$, $\mathcal{F}' \subseteq \Gamma$, under the assumption of feature independence, the OIQ of these two disjoint sets of features corresponds to the sum of the OIQs:*

$$\mathcal{I}_{\mathcal{F}\cup\mathcal{F}'}(d) = \mathcal{I}_{\mathcal{F}}(d) + \mathcal{I}_{\mathcal{F}'}(d) \ .$$

[2]For instance, just considering the occurrence of a few words as text features is enough to obtain an empty result in a standard web search engine.

In other words, under the assumption of independence between two feature sets, the resulting OIQ corresponds to the sum of their OIQs. Note that the feature independence assumption implies that the probability of documents improving projections in terms of several features is equivalent to the product of the probabilities over single features:

$$P\big(\{d' : \pi_{d'}(\gamma) \geq \pi_d(\gamma), \forall\ \gamma \in \Gamma\}\big) = \prod_{\gamma \in \Gamma} P\big(\{d' : \pi_{d'}(\gamma) \geq \pi_d(\gamma)\}\big)\ .$$

Strictly speaking, calculating the OIQ under this assumption sacrifices *dependency*, but the dependency within each feature subset can be preserved. In any case, our main goal is to simultaneously preserve specificity and quantitativity, which is not achieved by any existing representation method (see Table 2.1).

The second aspect to address when estimating the OIQ of a document is the document length. Our formalization starts from a set of potential documents $\mathcal{D}$. For homogeneity, this set is assumed to consist of sentences, or paragraphs, or posts, or full documents, etc. The problem is that the longer the documents are, the more challenging the cumulative probability estimation is. One way of managing long passages of text is by assuming *information additivity.*

Definition 3.6. Information Additivity: *Consider a document d, that is a concatenation of a set of document pieces, $\{\omega_1, \ldots, \omega_n\}$. Under the assumption of information additivity, the OIQ is the sum of the OIQs of each document piece:*

$$\mathcal{I}_\Gamma\big(d\big) = \sum_{i=1}^{n} \mathcal{I}_\Gamma\big(\omega_i\big)\ .$$

For instance, suppose that we consider words as document pieces, each of which is characterized by the word itself. Then, the OIQ of a concatenation of m words, $w_1, \ldots, w_m$, is expressed as:

$$\mathcal{I}\big(\{w_1, \ldots, w_m\}\big) = -\sum_{i=1}^{m} \log\big(P(w_i)\big)\ .$$

Note that considering words as document pieces is not equivalent to considering words as features due to the effect of repeated words:

$$\mathcal{I}\big(\{w_1, w_2, w_2\}\big) = -log\Big(P\big(\{d : tf(d, w_1) \geq 1\}\big) \cdot P\big(\{d : tf(d, w_2) \geq 2\}\big)\Big) \neq$$
$$\neq \mathcal{I}_\Gamma\big(\{w_1\}\big) + \mathcal{I}_\Gamma\big(\{w_2\}\big) + \mathcal{I}_\Gamma\big(\{w_2\}\big).$$

The accumulation of highly probable features can produce noise in the OIQ computation. The effectiveness of stopword removal or vocabulary reduction via the *tf·idf* criterion has been reported repeatedly in the literature. Supported by this principle, we can similarly apply feature projection reduction by truncating projections under a specified

OIQ threshold, which is denoted by th:

$$\mathcal{I}_{\Gamma}^{th}(d) = \begin{cases} \mathcal{I}_{\Gamma}(d), & \text{if } \mathcal{I}_{\Gamma}(d) \geq th \\ 0 & \text{in other case} \end{cases} . \tag{3.1}$$

In summary, given a partition of features such that $\Gamma = \Gamma^1 \cup \cdots \cup \Gamma^k$ and an OIQ threshold th, under the assumptions of feature independence and information additivity, the OIQ of a document that consists of a set of atomic text pieces $\{\omega_1, \ldots, \omega_m\}$ can be estimated as follows:

$$\mathcal{I}_{\Gamma}^{th}(\{\omega_1, \ldots, \omega_m\}) = \sum_{i=1}^{m} \sum_{j=1}^{k} \mathcal{I}_{\Gamma^j}^{th}(\omega_i) . \tag{3.2}$$

3.5. The OIQ vs. Other Text Representation Models

In this section, we relate the OIQ to several traditional methods of representing texts: the weighted vector space model, language models and distributional representations. In the following, to generalize binary and quantitative feature values, we define the notion of a feature indicator.

Definition 3.7. *Given a feature $\gamma \in \Gamma$, a feature indicator, denoted by χ_γ, is a mapping function from the set of documents to $\mathbb{R}$, defined as follows:*

$$\chi_\gamma(d) = \begin{cases} 1, & \textit{if } \gamma \in d \\ 0, & \textit{otherwise} \end{cases} .$$

The following properties state how the OIQ generalizes both the *idf* and *tf·idf* representation models:

Proposition 3.2. *Under the assumptions of word information additivity and equiprobability, the OIQ is equivalent to the tf representation. Let $d = (x_1, \ldots, x_n)$ be the tf representation of a document d with respect to the vocabulary (feature set) $\Gamma = \{\chi_{w_1}, \ldots, \chi_{w_n}\}$:*

$$\mathcal{I}_{\{\chi_{w_i}\}}(d) = tf(w_i, d) = x_i .$$

Proposition 3.3. *When word occurrences are taken as features, the OIQ of a word is equivalent to its* idf.

$$\mathcal{I}_{\{\chi_w\}}(w) = idf(w) .$$

Proposition 3.4. *When word occurrences are taken as features and under the assumption of information additivity, the OIQs of single features are equivalent to the* tf·idf *representation.*

$$\mathcal{I}_{\{\chi_{w_i}\}}(d) = tf(w_i, d) \cdot idf(w_i) .$$

Intuitively, *idf* corresponds to the OIQ of a word-feature, and *tf* is captured by the information additivity across words, that is, the OIQ that is associated with a word is accumulated across its appearances in the document.

In addition, language models can be generalized by considering word/position pairs as features:

Proposition 3.5. *When the occurrences of word-position pairs are taken as features and under the assumption that the documents in the collection are generated from a probability distribution Θ, the perplexity over the language model defined by θ of a word sequence $d = (w_1, \ldots, w_m)$ is an exponential function of the OIQ:*

$$Perplexity(d) = 2^{\frac{1}{m}\mathcal{I}_\Gamma(d)} \ .$$

Given the correspondence between $\mathcal{I}_\Gamma(w_1, \ldots, w_m)$ and $P(w_1, \ldots, w_m)$, we can infer the unigram model by assuming information additivity:

$$\mathcal{I}_\Gamma(w_1, \ldots, w_m) = -log\left(\prod_{i=1}^{m} P(w_i)\right) \ ,$$

or the n-gram model by considering the corresponding statistical assumptions:

$$\mathcal{I}_\Gamma(w_1, \ldots, w_m) = -log\left(\prod_{i=1}^{m} P(w_i \mid w_{i-(q-1)}, \ldots, w_{i-1})\right) \ ,$$

Traditional feature-set-based models and their information theory extensions are also generalized by the OIQ. For instance, Lin's distance can be expressed in terms of OIQs.

Proposition 3.6. *Given two documents d_1 and d_2, when word occurrences are taken as features and under the assumptions of feature independence and information additivity, Lin's distance can be expressed as:*

$$Lin(d_1, d_2) = \frac{\mathcal{I}(\mathcal{O}_\Gamma(d_1) \cap \mathcal{O}_\Gamma(d_2))}{\mathcal{I}(\mathcal{O}_\Gamma(d_1)) + \mathcal{I}(\mathcal{O}_\Gamma(d_2))} \ .$$

3.6. OIQ vs. Copulas and Information Algebra Theories

Copulas are models that describe the relationship between a multivariate distribution and the marginal distributions. For a random vector $\mathbf{X}$ with continuous marginal distributions $F_i(x_i)$, $i = 1, \ldots, d$, according to Sklar's theorem [88], a multivariate cumulative distribution can be expressed using the univariate marginal cumulative distributions and a copula function, C:

$$P(X_1 \leq x_1, \ldots, X_d \leq x_d) = F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \ .$$

Then, through the application of the probability integral transformation to the original random vector $\mathbf{X}$, it is possible to produce a uniform random vector $\mathbf{U} = (U_1, \ldots, U_d)$, with $U_i = F_i(X_i)$. The copula is the joint distribution function of the random variables

$U_i = F_i(X_i)$, $i = 1, \dots d$. In addition, given the joint distribution and the marginals it is possible to define a copula as:

$$P(U_1 \leq u_1, \dots, U_d \leq u_d) = C(u_1, \dots, u_d) = F(F^{-1}(u_1), \dots, F^{-1}(u_d)) \ ,$$

where the generalized inverse function G^{-1} of a cumulative density function G is defined as $G^{-1}(\mathbf{u}) := inf\{x \in \mathbb{R} \cup \{\infty\} : G(\mathbf{x}) \geq \mathbf{u}\}$ for all $\mathbf{u} \in (0, 1]$ and $G^{-1}(0) := sup\{x \in \mathbb{R} \cup \{\infty\} : G(\mathbf{x}) = 0\}$. A detailed treatment of copulas is given in [88, 62].

Copulas can facilitate the analysis of the structures of joint distributions. The OIQ can be expressed in terms of copulas, that is, cumulative joint distributions. Redefining each feature γ_i as a random variable whose distribution is determined by its projection values $\pi_d(\gamma_i)$ in $\mathcal{D}$, we obtain:

$$\mathcal{I}\big(X\big) = -log(\mathrm{C}(x_1, \dots, x_n)) = -log(P(\gamma_1 \geq x_1, \dots \gamma_n \geq x_n)) \ .$$

In this way, we can separately estimate each marginal distribution $F_i(.)$ and the dependency structure among the marginal distributions.

The copula is defined over $\leq$ comparisons. To adapt this structure to the OIQ, it is sufficient to either reverse the feature values or use a survival function. A multivariate survival function can be expressed using the univariate survival functions and the survival copula:

$$P(X_1 \geq x_1, \dots X_d \geq x_d) = \bar{F}(x_1, \dots, x_d) = \bar{C}(\bar{F}_1(x_1), \dots, \bar{F}_d(x_d)) \ .$$

Here, the survival copula is a distribution function on the hypercube (not a survival function) of the random vector $\bar{\mathbf{U}} = (\bar{U}_1, \dots, \bar{U}_d)$ with $\bar{U}_i = \bar{F}_i(X_i)$

$$P(\bar{U}_1 \leq u_1, \dots, \bar{U}_d \leq u_d) = \bar{C}(u_1, \dots, u_d) = \bar{F}(\bar{F}^{-1}(u_1), \dots, \bar{F}(\bar{F}^{-1}(u_d))) \ .$$

On the other hand, observational information theory has a connection with algebraic information theory [66, 65], which concerns the inclusion relationships between pieces of information. This theory refer to this relationship information as *order information* and the information quantity is interpreted as the set of questions that a document is able to answer. In contrast, in the OIQ, the information quantity of a document observation outcome is related to the number of representations across the universe of documents in which it is subsumed.

Like other information algebras, the OIQ permits operations between information representations. However, instead of operating directly on documents as in previous information algebras [65], the OIQ operates on fuzzy feature sets, without requiring the identity of the represented document to be specified.

3.7. Conclusions: The Generalization Power of the ORF and OIQ

In this chapter, we have proposed a representation framework that captures the three aspects analysed in the previous chapter: specificity, quantitativity and dependence. A review of these formal properties shows that the OIQ is invariant with respect to redundant features (dependence), monotonic with respect to the feature values (quantitativity) and sensitive to the likelihood of feature occurrences in the document collection (specificity). In particular, to our knowledge, it is the first representation model capable of capturing dependence, quantitativity and specificity simultaneously. The OIQ addresses these aspects by means of the following theoretical properties:

- While the probability sample spaces in previous representation frameworks have consisted of features (words, n-grams, etc.), in the ORF, the sample space is made up of the infinite and countable universe of documents. In other words, each document is a probabilistic event. Note that, with an infinite document space, this probability tends towards zero.
- To prevent the management of zero probability events, our framework distinguishes between a document itself and the corresponding *observation outcome*.
- Observation outcomes are modelled as feature fuzzy sets, capturing continuous values.
- The likelihood of an observation outcome is given by the mass probability of the documents subsuming it. We call this likelihood the Observational Information Quantity (OIQ).

We have validated the OIQ in terms of its generalization power. We have seen that the OIQ can be considered equivalent to most traditional representation methods, such as *idf*, *tf·idf*, and language models, under various statistical assumptions. The main challenge to be overcomed is OIQ estimation, which is related to copula theory in the case of quantitative features and to traditional dimensionality reduction techniques in the case of categorical (discrete) features. The assumptions of feature set independence and the information additivity are baseline practical solutions for OIQ estimation when integrating quantitative and binary feature sets and managing variable-length documents.

The above properties enable the application of union and intersection operators to document observation outcomes. Therefore, the OIQ can support similarity functions based on both feature sets and information theory. In the next chapter, we exploit this capability to review the theoretical foundations of similarity functions in the context of information access.

Part II

Similarity

4

Revisiting Similarity Axiomatic

4.1. Introduction

Computing similarity is a core problem that pervades, either implicitly or explicitly, many information access tasks. For instance, document retrieval systems are (at least partially) based on computing the similarity between queries and documents. Summarization, document clustering and many other text processing applications require computing the similarity between texts. Evaluation measures for text generation tasks (such as summarization or machine translation) compare the output of systems against reference models produced by humans. Beyond textual similarity, applications such as collaborative recommendation are based on estimating the similarity between users (based on their preferences and behaviour) and products (also based on user preferences).

In general, computing similarity requires, at a minimum, deciding how to represent items and how to compare representations. We focus on the second problem: finding general, suitable *similarity functions* able to operate accurately on as many kinds of items and representation models as possible. Widely used similarity functions include the cosine distance, the Euclidean distance, Jaccard distance, and Lin's similarity.

In this chapter, we analyse various similarity axiomatic and their suitability for information access scenarios. We will see that the properties of similarity functions and the representation framework are interdependent. In particular, the representation framework on which similarity functions are based determines which properties can be satisfied or even defined.

The Observational Representation Framework (ORF) presented in this thesis captures aspects of representations based on metric spaces, feature sets and information theory. This generality opens the door for the formal comparison of similarity functions under a unique theoretical framework. On the basis of the ORF, we define five general formal constraints for similarity functions. Four of these formal constraints can be synthesized into a single sufficient formal property called *similarity information monotonicity*. We will describe its relation to traditional theoretical similarity frameworks.

4.2. Previous Similarity Axiomatic Frameworks

In this section, we review the most popular similarity axiomatic frameworks. For each one, we examine boundary cases in which the axioms are at least debatable. For simplicity, in all our examples, the considered features are words; however, all our reasoning is equally valid for other features used to represent documents, such as n-grams, concepts, syntactic and semantic relationships, meta-data, user preferences in recommendation scenarios, or followers in a network.

Note that our word-based examples are simplifications, and the reader might argue that the problems found in previous axiomatic frameworks may be due to a poor or incomplete definition of the corresponding representational features. For instance, some word mismatch problems disappear if we use proper ontological concepts, instead of words, as features. Our focus, however, is on the similarity function as an abstract general mechanism for comparing documents, which should be equally valid regardless of the particular choice of features. In other words, our goal is to identify desirable properties of similarity functions in such a way that they do not need to be redefined for each representation model, task, or dataset.

4.2.1. Metric Spaces

The most traditional axiomatic framework originates from the concept of a metric space [106, 69]. In psychology, the assumption that similarity can be expressed as a distance in a metric space was proposed by Shepards. His purpose was to give a rational probabilistic argument for the origin of his *generalization law* [105]. In this theoretical framework, a document is projected into a multidimensional space $X = (x_1, x_2, \ldots, x_n)$ where x_i represents the projection of the document into the space with respect to feature γ_i. Then, the corresponding similarity axioms are directly inherited from the standard geometric distance axioms.

The first axiom is *maximality*, which states that every pair of identical objects achieves a maximal and constant similarity:

$$Sim(X, X) = Sim(Y, Y) \geq Sim(X, Y) \ .$$

Objections to maximality have already been raised in the context of cognitive science [68]. Based on experiments on several topics – such as the cognition of Morse code [103] and the cognition of rectangles varying in size and reflectance [16] – many researchers believe that the axiom of maximality is not consistent with human intuition. Tversky's experiments, in particular, showed that maximality (or minimality in distance) does not hold when a larger stimulus with more features is compared to a smaller stimulus with fewer features: if a stimulus shows more details, its level of perceived self-similarity increases [114]. Lee et al. [73] related this phenomenon to Hick's law: *the reaction time to a choice in a visual search becomes longer as the amount of information increases.*

An example of how maximality is violated in the context of information access problems

is query specificity in information retrieval. If a query is identical to a document, then the longer the query is, the more relevant the match. For instance, a two-word document is less relevant to an identical two-word query than a 200 word document is to an identical 200-word query. In both cases, we are faced with self-similarity, but in the second case, the object contains more information than in the first case, and the link between query and document has much more specificity.

The second axiom related to metric spaces is *triangular inequality*:

$$Sim(X, Y) \geq Sim(X, Z) + Sim(Z, Y) \ .$$

which has also been refuted in several cognitive experiments [102, 103]. Other studies also have found evidence against the third axiom, *symmetricity* [14, 115, 92]:

$$Sim(X, Y) = Sim(Y, X) \ .$$

From a cognitive point of view, the reason for discarding the symmetricity axiom is that, according to human perception, specific concepts tend to be closer to generic concepts than vice versa. For instance, Tversky found that subjects commonly perceived the concept of *"North Korea"* as being closer to *"Red China"* than vice versa. Such asymmetricity often plays a role in information access problems; for instance, query-document similarity in information retrieval is inherently asymmetric. In the context of natural language processing, Gawron found that an asymmetric lexical similarity measure, using parameters favouring the less frequent words, greatly improved the performance of a dependency-based vector model [46]. Other authors have also achieved improvements in the context of information access when modelling asymmetricity in similarity functions in multiple domains [45, 50, 91, 52, 60, 38]. Therefore, a straightforward symmetricity axiom is too restrictive for a general understanding of similarity. On the other hand, similarities in opposite directions are rarely compared in the context of information access systems; for instance, most test collections designed to test systems that compute semantic textual similarity are symmetric by design.

4.2.2. Tversky and Gati

Other axiomatics start from a featural representation of the form $X = \{\gamma_1, \ldots, \gamma_n\}$. In particular, Tversky and Gati [115] attempted to formulate axiomatics for similarity from an ordinal perspective, defining a *monotone proximity structure* that is based on three properties. The first property is *dominance*, which states that replacing features with shared features increases the similarity:

$$Sim(\{\gamma_1, \gamma_1'\}, \{\gamma_2, \gamma_2'\}) < min\{Sim(\{\gamma_1, \gamma_1'\}, \{\gamma_1, \gamma_2'\}), Sim(\{\gamma_1, \gamma_1'\}, \{\gamma_2, \gamma_1'\})\} \ .$$

When illustrated with words as features, this implies that the proximity of "*brown monkey*" to "*red cross*" is lower than that to "*brown cross*" because the feature "*red*" in the second case has been replaced with the common feature "*brown*". This axiom, however,

is grounded on the idea of independence across dimensions: in this example, dominance is assumed regardless of how "*red*" and "*brown*" interact with "*cross*". However, words – as well as other features – do not co-occur randomly. In the context of cognitive science, several authors have reported a lack of feature independence in similarity [49, 82] [1]. For instance, let us consider a clustering scenario in which images are grouped in accordance with their descriptors. Suppose that three images are tagged as *"Disney mouse"*, *"Mickey game"* and *"Mouse game"*. *"Mickey"* is commonly associated with the Disney character, while *"mouse game"* can be associated with other contexts. Therefore, even though they do not share any words, *"Disney mouse"* could be considered closer to *"Mickey game"* than to *"Mouse game"*, as expressed in Example 4.1 below, which contradicts the dominance axiom.

Example 4.1.

$$Sim(\text{"Disney } \textbf{mouse}\text{"}, \text{"Mickey game"}) > Sim(\text{"Disney } \textbf{mouse}\text{"}, \text{"}\textbf{Mouse}\text{ game"}) \ .$$

The second axiom is *consistency*, which states that the ordinal relation between similarities in one dimension is independent of any other dimension.

$$Sim(\{\gamma_1, \gamma_1'\}, \{\gamma_2, \gamma_1'\}) < Sim(\{\gamma_1, \gamma_1'\}, \{\gamma_4, \gamma_1'\}) \Leftrightarrow$$
$$Sim(\{\gamma_1, \gamma_2'\}, \{\gamma_2, \gamma_2'\}) < Sim(\{\gamma_1, \gamma_2'\}, \{\gamma_4, \gamma_2'\}) \ .$$

Again, this axiom is grounded on the assumption that the features are mutually independent, and we can find counterexamples in the context of textual similarity. For instance, the word *"mouse"* is closer to *"Mickey"* than to *"hardware"* in the context of Disney films, but this is not true in the context of computers and external devices (*"wireless"*). Therefore, one might consider the following relations.

Example 4.2.

$$Sim(\text{"}\textbf{Disney}\ \ \textit{mouse}\text{"}, \text{"}\textbf{Disney}\ \ \textit{Mickey}\text{"}) >$$
$$Sim(\text{"}\textbf{Disney}\ \ \textit{mouse}\text{"}, \text{"}\textbf{Disney}\ \ \textit{hardware}\text{"})$$
$$Sim(\text{"}\textbf{Wireless}\ \textit{mouse}\text{"}, \text{"}\textbf{Wireless}\ \ \textit{Mickey}\text{"}) <$$
$$Sim(\text{"}\textbf{Wireless}\ \textit{mouse}\text{"}, \text{"}\textbf{Wireless}\ \textit{hardware}\text{"}) \ .$$

Finally, the third constraint, *transitivity*, is grounded on a definition of *betweenness* that assumes the validity of *consistency* [115]. Therefore, Example 4.2 also contradicts this third axiom.

4.2.3. Feature Contrast Model

The best-known work of Tversky on similarity is the *Feature Contrast Model* [114]. Under the assumption that documents can be represented as sets of features, he defined

[1] Notice that cognitive studies manage a generic notion of *feature*. Here we exemplify with word features, but with the goal in mind that the principles apply to any type of feature

three basic axioms: *matching*, *monotonicity* and *independence*. Once more, all of them are based on the idea that the features are mutually independent. The *matching* axiom states that the similarity between two feature sets X and Y can be computed as a function of the intersection and difference of their feature sets:

$$Sim(X,Y) = \alpha \cdot f(X \cap Y) - \beta_1 \cdot f(X \setminus Y) - \beta_2 \cdot f(Y \setminus X) \ ,$$

where f represents a certain *salience* function. The second axiom, *monotonicity*, is closely related to *dominance*. It states that the similarity increases with an increase in the intersection between sets or decrease in their difference.

$$\left.\begin{array}{l}(X \setminus Y) \subseteq (X' \setminus Y') \\ (Y \setminus X) \subseteq (Y' \setminus X') \\ (X \cap Y) \supseteq (X' \cap Y')\end{array}\right\} \Longrightarrow Sim(X,Y) \geq Sim(X',Y') \ .$$

In addition, the similarity strictly grows if at least one of the inclusion relationships is strict. However, again, we know that this is not always true for texts, given that words (and text features in general) do not occur independently of each other. Indeed, adding different words to a pair of texts may increase their similarity, as in the following example where *"desktop"* and *"computer"* bring *"apple"* and *"mouse"*, respectively, into the context of computers.

Example 4.3.

$$Sim(\text{"Apple desktop"}, \text{"Mouse computer"}) > Sim(\text{"Apple"}, \text{"Mouse"}) \ .$$

Example 4.1 (from the previous section) also contradicts monotonicity, given that the similarity increases despite the fact that the intersection decreases and the difference increases.

The third property is *independence*. Its formalization is less intuitive than those of the other axioms (refer to [114] for a deep explanation). Essentially, it states that features affect similarity in an independent manner. However, as shown before, Example 4.2 contradicts independence. In addition, as stated above, previous work in the cognitive science literature raises objections to the assumption of independence [49, 82].

In the context of information access the advantage of considering statistical dependencies between features is corroborated by the use of multiple dimensionality reduction approaches at the representation level: latent semantic indexing, latent Dirichlet allocation, neural distributional representations such as Word2Vec [84] or BERT [37], etc. In this thesis, our purpose is to model dependency within a similarity axiomatic framework.

Note that a connection exists between the Feature Contrast Model and the metric-space-based axioms: the Feature Contrast Model satisfies the metric space axioms for certain parameter values. The following proposition is proved in the Appendix A.2 of this work.

Proposition 4.1. *The Feature Contrast Model satisfies the metric space axioms if* $\alpha = 0$ *and* $\beta_1 = \beta_2 > 0$ *(i.e.,* $Sim(X,Y) = -\beta \cdot f(X \setminus Y) - \beta \cdot f(Y \setminus X)$*) and the salience function is additive for disjoint feature sets:* $X \cap Y = \emptyset \Longrightarrow f(X \cup Y) = f(X) + f(Y)$.

In other words, the main contribution of Tversky's axioms with respect to metric spaces is that they consider the salience of common features in addition to differences, and its main limitation is the assumption of feature independence.

4.2.4. Similarity Axioms in Information Retrieval

Similarity plays a key role in information retrieval (IR). The most basic IR scenario can be interpreted as the problem of estimating the similarity between a query (which represents the needs of a user) and the documents in a collection. Fang and Zhai presented a seminal work on the axiomatic of information retrieval. These axiomatic, summarized below, have been used to improve search functions and term weighting models [42, 41, 27].

Considering both a document and a query consisting of multisets of words (with possible repeated words), $D = \{w_1, \ldots, w_n\}$ and $Q = \{w_1, \ldots, w_m\}$, respectively; the frequency $c(w, D)$ of words in document D; and any reasonable measure $td(w)$ for term discrimination (such as the inverse document frequency), the axioms of Fang and Zhai are expressed as follows.

- **TFC1:** A greater number of occurrences of a query term increases the document score:

$$\left.\begin{array}{l} Q = \{q\}, \quad |D_1| = |D_2| \\ c(q, D_1) > c(q, D_2) \end{array}\right\} \Longrightarrow S(Q, D_1) > S(Q, D_2) \ .$$

- **TFC2:** The increase in the score due to an increase in the frequency of a query term is smaller for larger frequencies:

$$\left.\begin{array}{l} Q = \{q\}, \quad |D_1| = |D_2| = |D_3| \\ c(q, D_3) = c(q, D_2) + 1 \\ c(q, D_2) = c(q, D_1) + 1 \end{array}\right\} \Longrightarrow S(Q, D_2) - S(Q, D_1) > S(Q, D_3) - S(Q, D_2).$$

- **TFC3:** Distinct query terms have a greater effect than repeated query terms:

$$\left.\begin{array}{l} Q = \{q_1, q_2\}, \quad |D_1| = |D_2| \\ c(q_1, D_1) = c(q_1, D_2) + c(q_2, D_2) \\ c(q_1, D_2) > 0, \quad c(q_2, D_2) > 0 \end{array}\right\} \Longrightarrow S(Q, D_2) > S(Q, D_1) \ .$$

- **TDC:** Discriminative terms have a greater effect:

$$\left.\begin{array}{l} Q = \{q_1, q_2\}, \quad |D_1| = |D_2| \\ c(q_1, D_1) = c(q_2, D_2) = 1 \\ c(q_2, D_1) = c(q_1, D_2) = 0 \\ td(q_2) > td(q_2) \end{array}\right\} \Longrightarrow S(Q, D_2) > S(Q, D_1) \ .$$

- **LNC1:** Non relevant terms decrease the score:

$$\left.\begin{array}{l} t \notin Q \\ c(t, D_1) = c(t, D_2) + 1 \\ \forall w \neq t \, (c(w, D_1) = c(w, D_2)) \end{array}\right\} \Longrightarrow S(Q, D_2) > S(Q, D_1) \ .$$

- **LNC2:** Concatenating the document with itself does not decrease the score:

$$\left.\begin{array}{l} q \in Q, \quad c(q, D_1) > 0 \\ \forall w \, (c(w, D_2) = k \cdot c(w, D_1)) \end{array}\right\} \Longrightarrow S(Q, D_2) \geq S(Q, D_1) \ .$$

- **TF-LNC:** The relevance score should not decrease with the addition of more query terms to the document:

$$\left.\begin{array}{l} Q = \{q\}, c(q, D_2) > c(q, D_1) \\ |D_2| = |D_1| + c(q, D_2) - c(q, D_1) \end{array}\right\} \Longrightarrow S(Q, D_2) > S(Q, D_1) \ .$$

These axioms can be interpreted as a refinement of Tversky's Feature Contrast Model in which the informativeness (discriminativeness) of features is taken into account. The axioms TFC1, TFC3, TF-LNC and LNC1 are closely related to Tversky's monotonocity axiom because they capture the effects of commonalities and differences between a query and a document. TDC is related to the informativeness of single features. TF2 and TF3 reflect the idea that the informativeness of features progressively decreases when they appear several times. Finally, LNC2 represents the idea that the informativeness of features must be normalized with respect to the document size.

As in the cases of Tversky's axioms and the metric space axioms, the main limitation of Fang and Zhai's framework is that it does not consider dependencies between features (although this is a practical assumption in the context of IR). In fact, these axioms do consider a restricted version of dependency because they assume that repeated features (words) are interdependent (in particular, a repeated feature contributes less information than different features; see axioms TF2 and TF3). However, this is not sufficient to capture the characteristics of Examples 4.1, 4.2 and 4.3.

With regard to Example 4.1, according to LNC1, removing *"Mickey"* should increase the score, as should adding *"mouse"*, according to TFC3. Therefore:

$$Sim(\text{"}Disney\ mouse\text{"}, \text{"}Mickey\ game\text{"}) < Sim(\text{"}Disney\ \textbf{mouse}\text{"}, \text{"}\textbf{Mouse}\ game\text{"}) \ .$$

With regard to Example 4.2, the above axioms do not state anything, although the extended constraints defined in [43] are sensitive to this situation. Moreover, this theoretical framework is oriented towards retrieval problems, so it deals only with cases in which different documents are compared against the same query. Therefore, Example 4.3 cannot be addressed by these axioms.

4.3. A Formal Similarity Constraint Set for Information Access

The analysis in the previous section suggests that similarity is still an unclear notion and that the basic constraints or axioms of similarity functions defined in IR and other research areas do not capture every relevant aspect in the context of information access. In general, they focus on how many features are shared and to what extent. They do not capture the specificity, informativeness and dependence of features, which are core elements in information theory. This shortcoming arises from the fact that the existing axiomatic and formal constraints are supported by geometric or set-based object representations. In this thesis, we take advantage of the expressive power of the proposed Observational Representation Framework (ORF), which generalizes vector- and feature-set-based representations while capturing the notion of specificity or unexpectedness from information theory. Based on this theoretical framework, we define a novel set of formal similarity constraints[2].

4.3.1. Notation: Representation and Similarity Functions

Here, we return to the notions of *document observation outcomes*, *occurrence* and the *Observational Information Quantity* (OIQ) (see Chapter 3). Since document observation outcomes are modelled as fuzzy feature sets, we can apply set operators to them. In addition, the membership values of a fuzzy feature set generalize the vector-based representations, and the probabilities of features can be captured via the notions of *occurrence* and the OIQ. Therefore, under this formal framework, we can analyse and compare similarity measures from different families.

Let $\mathcal{D}$ be a collection of documents, Γ be the universe of features and $\mathcal{P}(\Gamma)$ be the set of all possible fuzzy feature sets. We start by defining the similarity between fuzzy feature sets as follows.

Definition 4.1. *A similarity function,* $Sim : \mathcal{P}(\Gamma) \times \mathcal{P}(\Gamma) \rightarrow \mathbb{R}$*, is a mapping from* $\mathcal{P}(\Gamma) \times \mathcal{P}(\Gamma)$ *to the set of real numbers:*

That is, the input to a similarity function is any pair of fuzzy feature sets, and the output is a real value. Our axioms are defined in terms of larger/smaller similarity value comparisons. Therefore, the range of the similarity values (e.g., $(0..1)$, $(0..\infty)$, $(-\infty..\infty)$, etc.) is not relevant to our study.

The main hypothesis in this work is that there exists a universal set of basic similarity principles (formal constraints) that should be observed regardless of the feature space and the sample set of document representations. Therefore, our formal study does not prescribe these aspects and can accommodate various notions of similarity, which will depend on how the features and statistical events are defined.

[2]Note that we use the term "formal constraints" instead of “axioms” (which is more common in the literature) because they are meant to restrict the space of admissible similarity functions, rather than as a starting point for a deductive system (which is the usual interpretation of an axiomatics).

In the remainder of the chapter, for clarity, we will denote the concatenation of two disjoint fuzzy feature sets X and Y by XY:

$$XY \equiv (X \cup Y) \text{ where } X \cap Y = \emptyset \text{ and } X, Y \neq \emptyset \ .$$

In addition, for readability, we will denote the probability of occurrence of a fuzzy feature set. i.e., $P(Occ(X)) = P\big(\{d \in \mathcal{D} : \mathcal{O}_\Gamma(d) \supseteq X\}\big)$, simply by $P(X)$.

4.3.2. Formal Constraints

We define our formal constraints in terms of the behaviour of the similarity upon the concatenation of fuzzy feature sets as expressed above (XY). This is a simplified situation, as we do not consider aspects such as repeated words or quantitative features (such as temporal meta-data). However, we note that the more a set of constraints is able to characterize and discriminate similarity functions in such simple situations, the more powerful these constraints are.

Constraint 4.1. **[IDENTITY]:** *Adding or removing features to or from a fuzzy feature set decreases its similarity to the original set:*

$$Sim(X, X) > Sim(X, XY) \quad \textit{and}$$

$$Sim(XY, XY) > Sim(XY, X) \ .$$

This first constraint states that modifying a document representation (by removing or adding information) decreases its similarity to the original representation. Intuitively, *"if something changes, it is no longer the same"*. For instance, although we cannot axiomatically state how close *"Apple mouse"* is to itself, we can at least say that it is more similar to itself than to *"Apple"* or to *"Apple mouse desktop"*. This constraint is actually a relaxed version of maximality: we postulate that any document representation is more similar to itself than to any other representation, but not that its self-similarity is necessarily maximal. The reason to avoid the postulation of *maximality* is that, according to Tversky and many other authors (see Section 4.2.1), the more information an object contains, the more self-similar it is; in fact, this is our second constraint.

Constraint 4.2. **[IDENTITY-SP]:** *Adding new features to a fuzzy feature set increases its self-similarity:*

$$Sim(XY, XY) > Sim(X, X) \ .$$

This constraint matches the observations in previous works that specific features have a greater effect on similarity than generic features (see Section 4.2.1). Let us consider Figure 4.1. In the leftmost case, the red pair seems to be more similar than the rest, while when presenting the right distribution, the pair of white apples seems to be more similar. In both cases, the most similar apples are identical. The key point is that the less likely the documents are (or the more specific they are), the more they are similar to themselves.

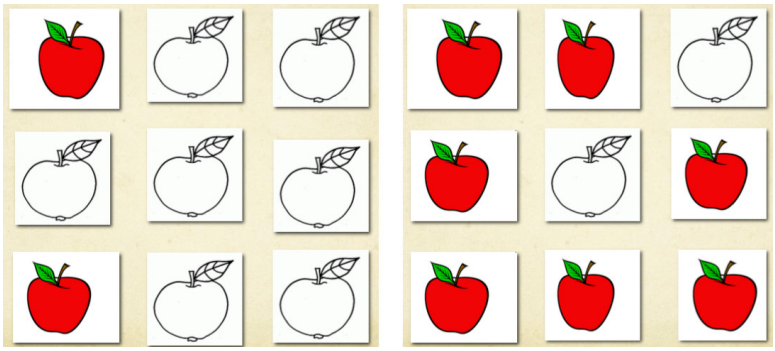

Figure 4.1: Red and white apples are considered the most similar object pairs by humans in the left and right side, respectively.

In the context of information retrieval, two documents with the same title are probably related, but two documents with the same content are certainly related. In other words, the more information two identical document representations contain (feature specificity), the more strongly we can ensure that the documents represented by those representations are similar. Let us illustrate this constraint with the following basic example.

Example 4.4.

$$Sim(\text{“}Apple\ mouse\ desktop\text{”}, \text{“}Apple\ mouse\ desktop\text{”}) > Sim(\text{“}Apple\ mouse\text{”}, \text{“}Apple\ mouse\text{”}) > Sim(\text{“}Apple\text{”}, \text{“}Apple\text{”}) \ .$$

Note that the second part of Constraint 4.1:

$$Sim(XY, XY) > Sim(XY, X) \ ,$$

can be directly derived from the first part:

$$Sim(X, Y) > Sim(X, XY) \ ,$$

together with Constraint 4.2:

$$Sim(XY, XY) > Sim(X, X) \quad \wedge \quad Sim(X, X) > Sim(X, XY) \implies \implies Sim(XY, XY) > Sim(X, XY) \ .$$

Therefore, we could reduce Constraint 4.1 to its first component. However, we maintain this partial redundancy so that our formal constraints will better align with the intuitive properties of similarity functions.

According to the literature, one limitation of Tversky's axioms is that they do not account for feature dependencies within the intersection and difference subsets [49]. To overcome this limitation, we define two constraints UNEXPECTEDNESS and DEPENDENCY that describe the expected behaviour of the similarity function in regard to feature dependencies (for a discussion in probabilistic terms, see Chapter 3).

Constraint 4.3. [UNEXPECTEDNESS]**:** *Adding unexpected features affects the similarity to a greater extent than adding expected features. Formally:*

$$\begin{aligned} &\textit{If } P(Y \mid X) < P(Y' \mid X) \textit{ then} \\ &Sim(X, XY) < Sim(X, XY') \ . \end{aligned}$$

For instance, suppose that we wish to improve the diversity of the search results for a query about "Mickey". Let us imagine that the first retrieval result is tagged as *"Mickey"*, and as the second result, we need to select between two images tagged as *"Mickey mouse"* and *"Mickey apple"*. *"Mickey mouse"* is more similar to *"Mickey"* than *"Mickey apple"* is, because *"apple"* is less likely to appear in the context of *"Mickey"* than *"mouse"* is. Therefore, we should select *"Mickey apple"* if we wish to improve the search result diversity.

Example 4.5.

$$Sim(\textit{"Mickey"}, \textit{"Mickey } \mathbf{mouse}\textit{"}) > Sim(\textit{"Mickey"}, \textit{"Mickey } \mathbf{apple}\textit{"}) \ .$$

Constraint 4.4. [DEPENDENCY]**:** *Adding new features to two fuzzy feature sets increases their mutual similarity if their respective conditional probabilities increase:*

$$\begin{aligned} &\textit{If } P(XZ \mid YZ') > P(X \mid Y) \quad \textit{and } P(YZ' \mid XZ) > P(Y \mid X) \\ &\textit{then } Sim(XZ, YZ') > Sim(X, Y) \ . \end{aligned}$$

This constraint introduces the possibility that adding different features to different document representations may bring them closer instead of necessarily making them less similar, as postulated by Tversky. For instance, in the following example, *"apple"* and *"mouse"* become closer to each other once the domain is specified by the additional words *"desktop"* and *"computer"*, respectively.

Example 4.6.

$$Sim(\textit{"Apple desktop"}, \textit{"Computer mouse"}) > Sim(\textit{"Apple"}, \textit{"Mouse"}) \ .$$

This example corresponds to the constraint through the following mapping: $X =$ *"apple"*, $Z =$ *"desktop"*, $Y =$ *"mouse"* and $Z' =$ *"computer"*. Thus, it is easier to find *"Computer mouse"* given the context *"Apple desktop"* than it is to find *"mouse"* given the context *"apple"* because *"apple"* is much more ambiguous. The constraint postulates, in other words, that this increase in similarity occurs when each of the two new pieces is more likely in the presence of the other.

Below, we prove that this formal constraint is not compatible with Tversky's framework; see Proof A.2.2.

Proposition 4.2. *Tversky's monotonicity axiom is not compatible with the* DEPENDENCY *constraint.*

As shown in the previous section, asymmetricity is an accepted characteristic of similarity. This assertion is corroborated by studies in both cognitive and computer science [14, 115, 92, 45, 50, 91, 52, 60, 38]. However, to our knowledge, the asymmetricity property and its behaviour have not yet been formalized. We attempt to formalize the concept of asymmetry with the following formal constraint.

Constraint 4.5. [ASYMMETRICITY]: *A fuzzy feature set is more similar to any of its parts than vice versa:*

$$Sim(XY, X) \geq Sim(X, XY) \ .$$

Intuitively, for instance, *Louis Armstrong* is more similar to the concept of *human* than vice versa, because *Louis Armstrong* has all the features of a human, whereas *human* describes Louis Armstrong in a very limited way. Another example, this time in the space of pieces of information, is the similarity of a medical-surgical book on human anatomy to an anatomical description of arms. The full book on human anatomy is more similar to the description of arms than vice versa because the first fully describes the second, but the converse is not true.

This constraint incorporates the results of multiple studies in cognitive sciences that have concluded that similarity is inherently asymmetric [114] because specific objects (feature supersets) are more similar to undefined objects (feature subsets) than vice versa. Under this constraint, the fuzzy feature set XY has more features and is therefore more specific than the fuzzy feature set X. In other words, XY describes X (and therefore, there is a strong directional connection), but X does not describe XY (and therefore the connection in this direction is weaker). The ASYMMETRICITY constraint is consistent with the salience imbalance hypothesis studied by Ortony in the context of cognitive science [92], which states that *"the salience of the ground (common features) is higher in the second component"*.

Note, however, that the notion of similarity in information access problems has certain unique characteristics, and there may be cases in which similarity is applied to tasks that are symmetric in nature. In a clustering task, for instance, there is no preferred direction for computing similarity, and therefore, the use of asymmetric measures can be superfluous or even counterproductive. Indeed, in practice, most reference datasets used to study the problem of similarity in information access are symmetric by construction. Hence, $\geq$ is used instead of $>$ in the constraint. We leave this last constraint as an open proposal for further discussion.

4.3.3. Similarity Information Monotonicity (SIM): A Sufficient Condition

We will now introduce a property called, *similarity information monotonicity* (SIM), which subsumes our first four constraints: any similarity function that complies with SIM also complies with IDENTITY, IDENTITY-SP, UNEXPECTEDNESS and DEPENDENCY. It may also comply ASYMMETRICITYbut does not necessarily do so. We will use SIM as the starting point to derive the similarity functions presented in this thesis. Note that

defining a sufficient condition for these four axioms makes it easier to check the formal suitability of similarity functions.

According to the analysis summarized in Table 5.1 (see following chapter), the pointwise mutual information and conditional probability together are able to satisfy our four basic formal constraints (only ASYMMETRICITY is violated by both the PMI and conditional probability). The intuition is that the PMI and conditional probability represent two complementary aspects of similarity, that might be combined into a single similarity measure that satisfies our four main constraints.

We now postulate the similarity information monotonicity axiom, in accordance with the above intuitions.

Definition 4.2. [**Similarity Information Monotonicity**] *If the PMI and conditional probabilities between two fuzzy feature sets grow, then the similarity between these sets also grows. Formally, if:*

$$\Delta PMI(X,Y) \geq 0 \quad \wedge \quad \Delta P(X \mid Y) \geq 0 \quad \wedge \quad \Delta P(Y \mid X) \geq 0 \ .$$

then $\Delta Sim(X,Y) \geq 0$. *In addition, if at least one increase is strict (*$>$*), then the increase in similarity is also strict.*

In other words, SIM basically states that the PMI and conditional probability are the two basic dimensions of similarity and that the similarity should be monotonic with respect to them. If both grow, then the similarity grows. In the case of a trade-off in the PMI and conditional probability values, SIM does not prescribe how the similarity behaves; it will depend on the particularities of the similarity function. SIM can be expressed in terms of increases in the joint and single information quantities, as stated in the following lemma; see Proof A.2.3.

Lemma 4.1. *SIM is equivalent to stating that a positive similarity increase occurs when both the information quantities of the compared fuzzy feature sets and their sum increase to a greater extent than the information quantity of the union of the fuzzy feature sets:*

$$\Delta\mathcal{I}(X) + \Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \Longleftrightarrow \Delta PMI(X,Y) \geq 0$$

$$\textit{and } \Delta\mathcal{I}(X) \geq \Delta\mathcal{I}(X \cup Y) \Longleftrightarrow \Delta P(X|Y) \geq 0$$

$$\textit{and } \Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \Longleftrightarrow \Delta P(Y|X) \geq 0 \ .$$

The most important aspect of SIM is that it is a sufficient condition for our four basic constraints. The following propositions state that satisfying SIM is a sufficient condition for satisfing the IDENTITY, IDENTITY-SP, DEPENDENCY and UNEXPECTEDNESS constraints; see Proof A.2.4.

Proposition 4.3. *Satisfying SIM is a sufficient condition to satisfy* IDENTITY, IDENTITY-SP, DEPENDENCY *and* UNEXPECTEDNESS *constraints.*

Given that SIM is defined in a symmetric manner, it cannot be a sufficient condition for the ASYMMETRICITY constraint. In fact, we instead have the following proposition; see Proof A.2.5.

Proposition 4.4. *SIM does not imply any constraint with respect to the ASYMMETRICITY conditions.*

Finally, although we have discarded Tversky's axioms due to the need to consider the dependencies between features, SIM has a direct correspondence with Tversky's monotonicity axiom under the assumption of feature independence:

Proposition 4.5. *Under the assumption of statistical independence between the intersection and difference components of two fuzzy feature sets and considering the OIQ as the salience function:*

$$\mathcal{I}(XY) = \mathcal{I}(X \cap Y) + \mathcal{I}(X \setminus Y) + \mathcal{I}(Y \setminus X) \ .$$

SIM is equivalent to Tversky's monotonicity axiom.

Let us summarize the properties of SIM: (i) it can be used to provide a single proof for all of our four main constraints (identity, identity specificity, unexpectedness and dependence); (ii) it models the traditional PMI and conditional probability as complementary components of similarity; and (iii) it has a direct correspondence with Tversky's axioms when independence is assumed between the intersection and difference components.

4.4. Conclusions: The Gaps in Previous Similarity Axiomatic

The importance of feature specificity and dependence in determining similarity has been reported by many authors in cognitive studies and in the context of information systems. However, according to our analysis, traditional axiomatic similarity frameworks fail to capture a number of intuitive characteristics in the context of information access tasks. Specificity is not captured by metric space similarity axioms, and dependency is not captured by Tversky's feature-set-based axioms (only partially at the representation level). In addition, geometric models assume feature independence, which is not consistent with information system scenarios.

The representation framework proposed in this thesis, the ORF, allows us to axiomatize the specificity and dependence of features in similarity functions. In particular, our five proposed basic constraints (identity, identity specificity, unexpectedness, dependency and asymmetricity) capture specific situations that are consistent with psychological studies and are not captured by previous axiom sets. In the next chapter we will study existing similarity functions under these constraints.

Finally, four of these five constraints can be synthesized into a single property (SIM). In addition, under certain statistical assumptions, SIM is equivalent to Tversky's monotonicity axiom. SIM examines the behaviour of the pointwise mutual information (PMI)

and conditional probability. More precisely, SIM states that increases in the PMI and the conditional probabilities in both directions between two objects must be accompanied by an increase in similarity. This formal result suggests that the PMI and conditional probability are the two main components of similarity [7]. In the next chapter, we will use this property to define a similarity function that satisfies all constraints.

5

Analysing Similarity Functions

5.1. Introduction

As seen in Chapter 3, the representation framework presented in this thesis, the ORF, allows us to capture notions from various representation paradigms. Under the ORF, the formal constraints proposed in Chapter 4 enable an analytical study of existing similarity functions from a novel perspective, capturing notions of representations based on feature sets, distance measures and information theory. In this chapter, we provide a global overview of existing similarity functions within a general theoretical framework. More concretely, we classify existing similarity functions into families according to their representation paradigms. We will see that the feature representation determines the formal properties of similarity functions. In the literature, the formal limitations of similarity functions are mitigated via feature selection and smoothing techniques. We will see that these drawbacks can instead be addressed by the similarity function itself.

Based on the ORF paradigm and the SIM property described in the previous chapter, we define the Information Contrast Model (ICM), a similarity function that generalizes the pointwise mutual information (PMI) and Tversky's linear contrast model. Consequently, the ORF connects information theory with feature-set-based similarity functions. We will also prove that among the similarity functions analysed in this thesis, the ICM is the only one that satisfies the similarity axioms defined in Chapter 4. Finally, we will present a study case consisting of estimating the similarities for the axiom counterexamples shown in the previous chapter. This study supports the ICM as a suitable theoretically grounded similarity function.

5.2. Similarity Functions from a Representational Perspective

We start with the categorization of the existing similarity functions proposed in [73]: *geometric*, *featural* and *alignment-based* functions, and we extend this categorization with three additional categories. Table 5.1 summarizes the properties satisfied by different

similarity functions, which will be discussed in the remainder of this section. We do not analyse transformational approaches, as they do not follow a particular representation scheme described in Chapter 2.

5.2.1. Similarity as a Distance in a Metric Space

As seen in Section 2.2, objects can be represented as points in a metric space, and many similarity functions are based on this paradigm, which satisfies the maximality, triangular inequality and symmetricity axioms. Typically, documents are modelled as vectors of word frequencies, and similarity is defined in terms of metric distances such as the Euclidean or cosine distance.

The IDENTITY constraint is satisfied by the metric space axiomatic, considering that IDENTITY is actually a relaxed version of maximality. However, maximality is not compatible with IDENTITY-SP: in a metric space, every document is maximally similar to itself regardless of its specificity. Therefore, the specificity of features is not considered (see line 1 in Table 5.1).

In practice, this drawback is mitigated at the representation level. Weighting mechanisms, such as the popular tf·idf, fulfil this role by assigning greater weights to infrequent words. Note that the second component $\left(\log\left(\frac{1}{p(w)}\right)\right)$ of the tf·idf expression has a correspondence with Shannon's information quantity [3]. Stopword removal is also a way of discarding frequent (non-informative) features.

According to the literature, the metric-space-based cosine similarity outperforms other measures, such as the Euclidean distance, in the context of information retrieval tasks. One possible reason is that the cosine similarity smooths the effect of highly frequent features by considering only proportionality instead of absolute differences between objects. In other words, the cosine similarity does not reward features if they are salient in both pieces of information. This property mitigates the lack of IDENTITY-SP. For instance:

$$Cosine((2,10),(1,12)) = Cosine((2000,10),(1000,12)) \ .$$

UNEXPECTEDNESS and DEPENDENCY are not satisfied by metric distances, considering that feature dependency is not directly captured by such a similarity function. In the context of textual similarity, many approaches mitigate this weakness by enriching text representations with semantic features from ontologies such as WordNet. Term dependencies can then be captured by means of relationships such as synonymy and hypernymy [83, 98].

UNEXPECTEDNESS and DEPENDENCY can also be approached by means of dimensionality reduction techniques, which avoid redundant features. One of the earliest approaches to be developed was latent semantic indexing (LSI) [34]. Later, in the 2000s, generative topic models such as latent Dirichlet allocation (LDA) [18] became popular. More recently, word embedding representations have gained popularity [85, 95]. Some word embedding techniques, such as PMI matrix factorization or skip-gram with negative

Table 5.1: Measure families and constraint satisfaction

Similarity function	IDENTITY	IDENTITY-SP	UNEXPECTEDNESS	DEPENDENCY	ASYMMETRICITY
	(X,X)>(X,XY) (XY,XY)>(XY,X)	(XY,XY)>(X,X)	(X,XY)>(X,XZ) if p(Y|X)>p(Z|X)	(XZ,YY')>(X,Y) if p(XZ|YY')>p(X|Y) and p(YY'|XZ)>p(Y|X)	(XY,X)> (X,XY)
Similarity as Distance in Metric Spaces					
(1) Euclidean, Cosine, Dice	✔	✗	✗	✗	✗
(2) Cosine plus Word2Vec	✔	✗	✔	✗	✗
Similarity as a Feature Set Operator					
(3) Linear Contrast Model (LCM)	✔	✔	✗	✗	✔
(4) Ratio Constrast Model (RCM) (Jaccard, Dice, precision and recall)	✔	✗	✗	✗	✔
Similarity as Information-Theoretic Operator					
(5) Lin's model	✔	✗	✔	✗	✔
(6) String Lin's distance	✔	✗	✗	✗	✗
(7) Residual Entropy Similarity	✔	✔	✔	✗	✔
Similarity as a Comparison of Probabilistic Density Functions					
(8) Kullback-Leibler	✔	✗	✗	✗	✗
Similarity as a Probabilistic Generative Process					
(9) n-grams Language Model	✔	✔	✗	✗	✔
Similarity as a Probabilistic Event Operator					
(10) Conditional Probability	✔	✗	✔	✔	✗
(11) Pointwise Mutual Information	✔	✔	✗	✔	✗

sampling (Word2Vec), allow the estimation of probabilistic similarity functions. More specifically, a correspondence exists between the scalar product of two vectors and their pointwise mutual information (PMI) [74, 11]. As we will analyse in Section 5.2.6, the PMI satisfies IDENTITY-SP and DEPENDENCY but not UNEXPECTEDNESS. In addition, the cosine similarity under the Word2Vec representation, which has been used as a baseline in many tasks, satisfies UNEXPECTEDNESS(see Proof A.3.1) but neither IDENTITY-SP nor DEPENDENCY. Given that the cosine distance is a metric distance, we can guarantee that it does not satisfy IDENTITY-SP. However, we are unable to check analytically check whether it satisfies DEPENDENCY. We leave this analysis for future work.

Proposition 5.1. *Under the assumption that the PMI of words is not negative, the cosine similarity under the skip-gram with negative sampling representation satisfies UNEXPECTEDNESS.*

There are alternative approaches that additionally consider the user context. In the context of IR, Fuhr et al [44] proposed a method in which documents are represented in terms of their relevance to different queries (according to a standard IR function). In addition to applying user-oriented dimensionality reduction, this approach mitigates non-compliance with UNEXPECTEDNESS, as features appearing in all queries will not affect the similarity. It also mitigates non-compliance with the DEPENDENCY constraint, as redundant features will reinforce the similarity of documents to identical queries.

Overall, the main conclusion that we can draw from this analysis is that the recent improvements to metric space distances have been focused on indirect ways of satisfying UNEXPECTEDNESS and DEPENDENCY. Strictly speaking, IDENTITY-SP is an underlying shortcoming derived from the maximality axiom.

5.2.2. Similarity in the Form of Feature Set Operators

From this perspective, objects are represented as sets of features (see Section 2.3). Similarity is characterized based on Tversky's axioms (matching, monotonicity and independence). As explained in previous sections, a key contribution of Tversky is the *representation theorem*, which states that similarity can be modelled as a linear function of the intersection and difference of sets. This is known as Tversky's linear contrast model. Given two feature sets, X and Y :

$$Sim(X,Y) = \alpha_1 \cdot f(X \cap Y) - \beta_1 \cdot f(X \setminus Y) - \beta_2 \cdot f(Y \setminus X) \ ,$$

where f is a certain salience function that is monotonic with respect to set subsumption ($f(X) < f(X \cup Y)$).

As discussed in previous sections, Tversky's independence axiom is not compatible with UNEXPECTEDNESS or DEPENDENCY (see Proposition 4.2). However, its parametrization (β_1 and β_2) can capture ASYMMETRICITY. The linear contrast model also captures IDENTITY-SP, given that (see line 3 in Table 5.1):

$$Sim(X,X) = \alpha_1 \cdot f(X \cap X) - \alpha_2 \cdot f(X \setminus X) - \alpha_3 \cdot f(X \setminus X) =$$

$$= \alpha_1 \cdot f(X) - \alpha_2 \cdot f(\emptyset) - \alpha_3 \cdot f(\emptyset) = \alpha_1 \cdot f(X) \ .$$

As an alternative, Tversky proposed the ratio contrast model,

$$Sim(X,Y) = \frac{\alpha_1 \cdot f(X \cap Y)}{\alpha_2 \cdot f(X \setminus Y) + \alpha_3 \cdot f(Y \setminus X) + \alpha_4 \cdot f(X \cap Y)} \ ,$$

which is more commonly applied in the literature. In fact, many set-based similarity measures can be derived from the ratio contrast model by taking the set size as the salience function f. Maurice Ling reported a comprehensive description of these measures [78]. By fixing different values for $\langle \alpha_1, \alpha_2, \alpha_3, \alpha_4 \rangle$, we can obtain measures such as the Jaccard distance ($\langle 1,1,1,1 \rangle$), the precision ($\langle 1,1,0,1 \rangle$), the recall ($\langle 1,0,1,1 \rangle$), the Dice coefficient ($\langle 2,1,1,2 \rangle$), the Anderberg coefficient ($\langle 1,2,2,1 \rangle$) or the first Kulczynski coefficient ($\langle 1,1,1,0 \rangle$).

Tversky's empirical studies showed that the linear contrast model parametrization is highly sensitive to the particular scenario. One of the reasons why the ratio contrast model appears more commonly in the literature is its robustness to different parametrizations. More formally, the ratio contrast model satisfies the following proposition; see Proof A.3.2.

Proposition 5.2. *Whenever $\alpha_1 = \alpha_4$, any variation in α_1 and α_4 in the ratio contrast model produces ordinal equivalent similarity functions.*

In other words, only the relative values of α_2 and α_3 must be estimated to mantain a consistent ordering between similarity values. The drawback of the ratio formulation is that IDENTITY-SP is no longer satisfied; see Proof A.3.3 and line (4) in Table 5.1.

5.2.3. Similarity as an Information-Theoretic Operator

In the literature, feature-set-based similarity functions have been enriched with notions imported from information theory. Basically, such enrichment relies on considering features as probabilistic events. In this way, feature salience can be modelled in terms of information content (IC) and entropy.

For instance, in the context of cognitive science, Lee et al. [73] incorporated the IC into Tversky's featural model. Analogously, in the context of computer science, Lin's model extends Tversky's ratio contrast model by capturing the specificity of features in terms of Shannon's theory [77]. Given a set of formal assumptions, Lin obtained the *similarity theorem*, which is stated as follows.

Theorem 5.1 (Lin's Similarity Theorem). *The similarity between X and Y is measured by the ratio between the amount of information needed to express the commonality of X and Y and the information needed to fully describe what X and Y are:*

$$Sim(X,Y) = \frac{\log P(common(X,Y))}{\log P(description(X,Y))} \ ,$$

where $common(X, Y)$ *is a proposition that expresses the commonalities between* X *and* Y*, and* $description(X, Y)$ *is a proposition that describes what* X *and* Y *are.*

For text string similarity, Lin instantiated the model into the following similarity measure, considering words as independent features:

$$Lin(X, Y) = \frac{2 \cdot \sum_{w \in X \cap Y} I(w)}{\sum_{w \in X} I(w) + \sum_{w' \in Y} I(w')} \ .$$

This expression is consistent with Tversky's ratio contrast model, if we use the IC $(-\log P(X))$ as the f function [21].

Assumption 4 (maximality) in Lin's work [77], intrinsically contradicts the IDENTITY-SP constraint. DEPENDENCY also cannot be satisfied, as adding features necessarily increases the information in the denominator. However, Lin's model does capture UNEXPECTEDNESS, as the addition of unexpected features increases the denominator to a greater extent. Unfortunately, this property is lost under the assumption of independence in Lin's measure for text string similarity (see lines (5) and (6) in Table 5.1).

Cazzanti and Gupta [21] attempted to improve Lin's similarity by applying the linear contrast model with fixed parameters instead of the ratio contrast model, obtaining the Residual Entropy Similarity (RES):

$$RES = f(X \cap Y) - 0.5 \cdot f(X \setminus Y) - 0.5 \cdot f(Y \setminus X) \ ,$$

where the salience function f is the conditional entropy of random pieces of information, R, with respect to the observed features:

$$f(X) = H(R \mid X \subset R) \ .$$

Essentially, this salience function assigns more weight to infrequent features. The main contribution of the RES with respect to Lin's measure in terms of our axioms is compliance with IDENTITY-SP. However, it has the same limitations as Lin's model in terms of DEPENDENCY. More explicitly, the RES satisfies Tversky's monotonicity axiom (Property 8 in [21]) which is not compatible with DEPENDENCY (see line (7) in Table 5.1). Regarding ASYMMETRICITY, these measures take fixed parameters that make them symmetric; however, the parameters can be fixed to induce asymmetricity in the direction required by the constraint.

In conclusion, set-based similarity functions can be extended to probabilistic events to capture IDENTITY-SP and UNEXPECTEDNESS. The underlying limitation of both the RES and Lin's measures is that the intersection and difference sets are managed independently. The effect is that feature dependence across these subsets is ignored, meaning that DEPENDENCY cannot be satisfied (see lines (5), (6) and (7) in Table 5.1).

5.2.4. Similarity as a Comparison of Probability Distributions

Objects can be represented as probability distributions of features. Cha et al. described 65 different similarity measures based on comparisons of probability density functions [22]. This perspective offers remarkable generalization power, and in fact, metric distances and set-based similarity functions can be interpreted in these terms [22]. However, a common limitation of this perspective is that the corresponding measures do not satisfy UNEXPECTEDNESS and DEPENDENCY. The reason is that representing documents as independent probability distributions does not allow statistical dependencies to be inferred from a document collection.

In addition, none of these measures complies with IDENTITY-SP. The reason is that in this approach, a distribution is equally similar to itself regardless of how much information it contains. Even measures based on Shannon's entropy [22] assign a maximal similarity (or minimal distance) to identical distributions. Consider the most representative measure, the Kullback-Leibler divergence. Let P_i and Q_i denote the probabilities of feature i in the pieces of information P and Q, respectively; their divergence is:

$$d_{kl} \equiv \sum_i P_i \ln \frac{P_i}{Q_i} \ .$$

If $P_i = Q_i$ for all i, then, $d_{kl} = \sum_i P_i \ln 1 = 0$. Therefore, every object is equally similar to itself. The same behaviour is found with other distribution-entropy-based measures, such as the Jeffreys divergence, the K-divergence and the Jensen-Shannon divergence.

In addition, modelling objects as probability distributions does not allow the probabilities of objects to be modelled. For this reason, UNEXPECTEDNESS and DEPENDENCY cannot be satisfied. Finally, at least in the case of the Kullback-Leibler divergence, ASYMMETRICITY is not formally satisfied. Zero probability features in either distribution are not considered in either direction (see line (8) in Table 5.1).

5.2.5. Similarity as a Probabilistic Generative Process

Another way of modelling similarity is via the likelihood of a feature sequence given a probability distribution. This is the case for perplexity in language models (see Section 2.5). For instance, in the information retrieval approach proposed by Ponte and Croft [97], the similarity between a query q and a document d is estimated as the probability of the query being produced from the probabilistic distribution θ_d inferred from the document, $(Sim(q, d) = p(q \mid \theta_d))$. In particular, their approach assumes that θ_d is a multiple Bernoulli distribution:

$$p(q \mid \theta_d) = \prod_{w \in q} p(w \mid d) \cdot \prod_{w \notin q} \big(1 - p(w \mid d)\big) \ ,$$

where $p(w \mid d)$ is estimated as the relative frequency of the word w in the document. In practice, this requires a smoothing process in which the probabilities of unseen query words are estimated from the whole collection. Many improvements have subsequently

been proposed. For instance, Hiemstra and Kraaij [57] and Miller et al. [86] proposed a variation based on multinomial word distributions.

In general, language models can satisfy IDENTITY and IDENTITY-SP. For instance, the last component in the model proposed by Zhai and Lafferty [124] is the sum of the probabilities of the query terms in the collection ($\ldots + \sum_{w \in q} p(w \mid \mathcal{D})$). This component is not considered in a document retrieval task because it does not affect the document ranking, but it will increase the self-similarity of larger queries, as our constraint requires.

Strictly speaking, UNEXPECTEDNESS is not satisfied, as it is not possible to estimate the dependency between unseen query words and the document because the probability distribution is inferred from the document itself. However, according to Zhai's analysis [123, 124], smoothing techniques have a correspondence to the idf effect, and therefore mitigate non-compliance with our UNEXPECTEDNESS constraint. Similar to the case of probability-distribution-based measures, DEPENDENCY cannot be satisfied because there is no probabilistic space external to both representations (see line (9) in Table 5.1). This limitation is partially mitigated by the use of n-grams, which capture the dependencies of some word-based features. Finally, ASYMMETRICITY is satisfied. Note that the perplexity of a word sequence given a distribution inferred from a document containing that sequence is lower than the perplexity of the full document given the distribution inferred from the sequence.

5.2.6. Similarity as a Probabilistic Event Operator

From this perspective, pieces of information are represented as compound events in an overall probabilistic distribution; see Section 2.6. Their similarity is determined from the distribution of features (single events). This paradigm captures psychological notions such as specificity and the diagnosticity principle: "features are more salient if they help discriminate objects".

From this perspective, similarity can be modelled as a conditional probability of feature sets:

$$Sim(X, Y) = P(X \mid Y) \ .$$

The strength of conditional probability as a similarity function is that it trivially satisfies DEPENDENCY and UNEXPECTEDNESS. Adding different features to the second piece of information can increase the estimated similarity. For instance:

$$P\big(\text{“Computer”} \mid \text{“Apple Desktop”}\big) > P\big(\text{“Computer”} \mid \text{“Apple”}\big) \ .$$

The main weakness of conditional probability is IDENTITY-SP. The self-similarity is maximal and constant for any pair of subsumed pieces of information; see line (10) in Table 5.1:

$$X \subseteq Y \Longrightarrow P(X \mid Y) = 1 \ .$$

.

An alternative, and one of the most popular similarity functions developed from this perspective, is the pointwise mutual information (PMI), which is based on the idea that the more highly statistically correlated two pieces of information are, the more similar they are. The PMI is computed as:

$$PMI = \log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right) ,$$

which is equal to zero when the two events are independent. The PMI has been used in multiple approaches to estimate pairwise word similarity. As in the case of conditional probability, we can prove the following; see Proof A.3.4.

Proposition 5.3. *The PMI satisfies* DEPENDENCY.

From the point of view of our constraint framework, the main strength of the PMI is that, unlike other functions, it complies with the IDENTITY-SP constraint. In particular, the self-similarity of any piece of information corresponds to its *information quantity*:

$$PMI(X,X) = \log\left(\frac{P(X,X)}{P(X)\cdot P(X)}\right) = -\log(P(X)) .$$

The main shortcoming of the PMI is that (as its name suggests) it focuses only on common features. For this reason, the following proposition holds; see Proof A.3.5 and line (11) in Table 5.1.

Proposition 5.4. *The PMI does not satisfy* UNEXPECTEDNESS.

5.2.7. Summary of the Theoretical Analysis of Similarity Functions

As noted previously, the effect of feature specificity on similarity has been reported by many authors in cognitive studies, in the form of principles such as the diagnosticity principle [114] and the size principle [111]. Analogously, in the context of information systems, most similarity approaches, such as the idf feature weighting, the Information Content in Lin's similarity function and the residual entropy similarity, or perplexity and smoothing in language models, capture feature specificity in some way. All of them have a direct connection with Shannon's information quantity, which has been considered a core concept of similarity since the earliest studies, such as Hick's work in 1952 [56].

In addition, feature dependency has been at least partially captured at the representation level; for instance, by means of dimensionality reduction in metric spaces, ontologies, or n-grams in language models.

Both feature specificity and feature dependence must be captured to comply with the IDENTITY-SP, UNEXPECTEDNESS and DEPENDENCY constraints. Table 5.1 summarizes the properties satisfied by particular similarity functions organized by families. As the table shows, the property most frequently missing is DEPENDENCY: it is not captured by

transformational, feature- or event-set-based, density-function-based or language model approaches.

To capture DEPENDENCY as a property of the similarity function (i.e., regardless of the representation level), we must look at models that represent documents as compound probabilistic events. In this familyof models, we find the PMI and the conditional probability, which have complementary properties. Both comply with DEPENDENCY, but the PMI does not satisfy UNEXPECTEDNESS, whereas the conditional probability does not comply with IDENTITY-SP. This observation suggests that these two functions could serve complementary purposes in defining a theoretically grounded similarity function. In fact, in later sections, we will see that our proposed similarity function maps to either the PMI or the conditional probability when its parameters are set at the external boundaries of the range allowed by the formal constraints.

5.3. Proposed Similarity Function: The Information Contrast Model (ICM)

Taking the SIM property as the core requirement for similarity measures, we now derive a new similarity function, the *Information Contrast Model* (ICM)[1]. The SIM property suggests that a similarity function should consider the relative increases in the individual, sum and union information quantities. That is, $\mathcal{I}(X)$ and $\mathcal{I}(Y)$ should have a positive effect on similarity while $\mathcal{I}(X \cup Y)$ should have a negative effect.

ICM is defined over the notion of Observation Outcome (See Chapter 3), that is, feature fuzzy sets. Recall that this representation generalizes both feature vectors and feature sets. In addition, the notion of observation outcome has an associated probability via its occurrence set which was defined as:

$$Occ(X) = \left\{ d \in \mathcal{D} : \mathcal{O}_\Gamma(d) \supseteq X \right\} .$$

The occurrence probability $P(Occ(X))$ of the observation outcome X is defined over the sample space of documents. As a result, there exists a formal correspondence between feature sets (observation outcomes) and probabilistic events in such a way that:

$$P(Occ(X), Occ(Y)) = P(Occ(X) \cap Occ(Y)) = P(Occ(X \cup Y))$$

and therefore $\mathcal{I}(X, Y) = \mathcal{I}(X \cup Y)$.

Definition 5.1. *The Information Contrast Model value for a pair of observation outcomes is the linear combination of the observational information quantity of each fuzzy feature set and the observational information quantity of their union:*

$$ICM_{\alpha_1,\alpha_2,\beta}(X, Y) = \alpha_1 \cdot I(X) + \alpha_2 \cdot I(Y) - \beta \cdot I(X \cup Y) .$$

[1] We have selected this name by analogy with the Linear and Ratio Contrast model proposed by Tversky.

Under the Observational Representation Framework, this similarity function can be interpreted as a generalized parametric version of the pointwise mutual information, which is equivalent to

$$ICM_{\alpha_1,\alpha_2,\beta}(X,Y) = \log\left(\frac{P(Occ(X),Occ(Y))^{\beta}}{P(Occ(X))^{\alpha_1}\cdot P(Occ(Y))^{\alpha_2}}\right)$$

An interesting characteristic of the ICM relative to other similarity functions is that the feature sets are not dissected into their intersection or differences, thus allowing the feature dependencies to be captured appropriately. This is because the similarity is defined in terms of the information quantities of the individual feature sets and their union.

5.3.1. Formal Properties

The ICM satisfies SIM for a certain range of its parameters; see Proof A.3.6.

Proposition 5.5. *The ICM satisfies the similarity information monotonicity axiom when* $\alpha_1 + \alpha_2 > \beta > \alpha_1 > \alpha_2$.

The ICM will also satisfy ASYMMETRICITY if we add the restriction $\alpha_1 > \alpha_2$, which preferentially rewards the directional similarity from the more specific text to the more general text, as required by the ASYMMETRICITY constraint.

In a symmetric scenario, we can fix $\alpha_1 = \alpha_2 = 1$ without loss of generality; and then, the condition to satisfy SIM is simply $2 > \beta > 1$. With β in this range, the ICM satisfies all four basic constraints.

The ICM has a direct relationship with the pointwise mutual information and the product of conditional probabilities for certain values of its parameters (See Proof A.3.7):

Proposition 5.6. *The ICM generalizes the pointwise mutual information and the product of conditional probabilities.*

$$ICM_{\alpha_1=1,\alpha_2=1,\beta=1}(X,Y) = PMI(Occ(X),Occ(Y))$$

$$ICM_{\alpha_1=1,\alpha_2=1,\beta=2}(X,Y) = \log(P(Occ(X) \mid Occ(Y)) \cdot P(Occ(Y) \mid Occ(X))) \ .$$

Note that the PMI ($\beta = 1$) and the product of conditional probabilities ($\beta = 2$) are the (outer) limits of the ICM with the values allowed by the constraints ($1 < \beta < 2$). In other words, the PMI and conditional probability measures are extreme cases of the more general measure, the ICM.

The ICM is also closely related to set- and information-theory-based measures. It is a generalization of the linear contrast model:

Proposition 5.7. *If independence is assumed between the features in the intersection and difference subsets in the ICM and the information quantity is used as the salience*

function in the linear contrast model, then the ICM and the linear contrast model are equivalent.

$$ICM_{\alpha_1,\alpha_2,\beta}(X,Y) = (\alpha_1+\alpha_2-\beta)\cdot\mathcal{I}(X\cap Y)-(\beta-\alpha_1)\cdot(\mathcal{I}(X\setminus Y))-(\beta-\alpha_2)\cdot(\mathcal{I}(Y\setminus X)) \ .$$

5.4. Case Study: Capturing Counterexamples

Here, we present a basic empirical proof of concept of the ability of the ICM to capture the characteristics of the counterexamples used in Section 4.2 to invalidate, in the context of textual similarity, some of the axioms proposed in the literature.

Table 5.2: Case study: how ICM satisfies the intuitive inequalities used as examples along this paper. Each inequality illustrates an expected behavior that contradicts some axiom in the literature, or exemplifies the need for one of our formal constraints. Computing ICM using co-ocurrence statistics in Flickr, it complies with all the expected inequalities.

Intuitive inequality $Sim(A) > Sim(B)$	ICM_1	ICM_2	Satisfied?
Example for IDENTITY-SP / counterexample for maximality axiom			
Sim("apple computer","apple computer") > Sim("apple","apple")	1.32	0.88	✓
Example for ASYMMETRICITY / counterexample for symmetricity axiom			
Sim("north korea","china") > Sim("china","north korea")	2.86	-0.79	✓
Counterexample for dominance axiom			
Sim("disney mouse","game mickey") > Sim("disney **mouse**","game **mouse**")	1.32	0.88	✓
Counterexample for consistency and independency axioms (I)			
Sim("mouse **disney**","mickey **disney**") > Sim("mouse **disney**","hardware **disney**")	2.86	-0.79	✓
Counterexample for consistency and independency axioms (II)			
Sim("mouse **wireless**","hardware **wireless**") > Sim("mouse **wireless**","mickey **wireless**")	2.6	2.47	✓
Example for DEPENDENCY / counterexample for monotonicity Axiom			
Sim("apple desktop","mouse computer") > Sim("apple","mouse")	4.03	-2.86	✓
Example for IDENTITY axiom			
Sim("**apple mouse**","**apple mouse**") > Sim("apple **mouse**","**mouse**")	4.06	2.29	✓
Example for UNEXPECTEDNESS			
Sim("**mickey**","**mickey** mouse") > Sim("**mickey**","**mickey** apple")	2.59	1.51	✓

To do so, we need to estimate the information quantities of phrases such as *"Mickey Mouse"* or *"Apple desktop"*. We have used statistics from the Flickr search engine because it gives exact numbers. Web search engine statistics are much larger but offer only approximate statistics on the number of hits, and the numbers returned depend on the length of the query. For instance, for the word set *"Mickey apple"*, Flickr finds $2,141$ documents. Given that Flickr stores approximately $13,000$ million photos, this represents a probability of $0.164 \cdot 10^{-6}$. We performed this estimation for every text used in the examples and we computed the ICM value for each pair of texts. We set the ICM parameters to $\alpha_1 = 1.2, \alpha_2 = 1$, and $\beta = 1.5$, arbitrary values that lie in the

ranges specified in our theoretical analysis.

Table 5.2 shows the results. The first column presents the similarity inequality that we intuitively expect. The second and third columns present the ICM values of the leftmost and rightmost text pairs, respectively, in each inequality, and the last column indicates whether the ICM result agrees with our intuition. The first example, for instance, shows that the ICM assigns a higher self-similarity to "Apple computer" than to "Apple", in agreement with our identity specificity axiom and in disagreement with the maximality axiom from the literature. Overall, the ICM satisfies our axioms for all examples and violates other previous axioms in the cases in which they predict counterintuitive results.

This is, of course, anecdotal evidence rather than a quantitative confirmation of the validity of the ICM, but it serves as a proof-of-concept of how the ICM works in extreme situations derived from the formal analysis.

5.5. Conclusions: The Generalization Power of the ICM as a Similarity Function

Our analysis has shown that the typology of similarity functions determines their weaknesses from a formal perspective. Existing similarity functions can be categorized into a set of families: metric distances (e.g., cosine distance, Euclidean distance), operators over features or event sets (e.g. Jaccard similarity, Lin's similarity), measures of the proximity of probability density functions (e.g., Kullback-Leibler divergence), measures of the likelihood of pieces of information under probability distributions inferred from other pieces of information (e.g., language models) and probability distributions of pieces of information in a whole space (e.g., pointwise mutual information). None of the existing similarity functions known to us complies with the constraints proposed in Chapter 4. In most cases, this limitation is due to the nature of a similarity function in terms of its typology. Existing similarity approaches mitigate the limitations of their similarity functions in the feature selection stage. Some examples include tf·idf weighting, n-grams, and dimensionality reduction methods (LDA, word embeddings, etc).

Our analysis has also shown that the PMI and conditional probability are the two main components of similarity. First, the PMI and conditional probability together are able to satisfy all axioms directly at the similarity function level, but it is necessary to combine them to satisfy every axiom simultaneously. Second, the proposed constraints (except ASYMMETRICITY) can be derived from a single property, which we call *similarity information monotonicity*, in which the PMI and conditional probability are taken as the basic complementary components of similarity. Third, both can be generalized into a unique parametrizable function: the ICM.

We have also seen that a connection exists between traditional featural axiomatic (Tversky's model), metric spaces, and information-theoretic similarity functions (PMI and conditional probability). In particular, the ICM generalizes similarity functions from these families.

The main challenge identified in our study is how to properly estimate the information quantities (or probability of feature sets) for documents without assuming feature independence. Another open issue is whether the ASYMMETRICITY constraint might be helpful for information access tasks in general or is instead a property that holds only for problems that can be directly mapped to our cognitive understanding of language and concepts.

Note that our proposed similarity function, independently of the feature space, does not completely prescribe a unique notion of similarity; only its combination with a particular feature space (which should be linked to some practical scenario) leads to an operational notion of similarity. For instance, *Barcelona* may be closer to *Paris* or to *Madrid* depending on whether the focus is geopolitical issues or touristic sites, and this difference should be reflected in how features are distributed in the document collection of interest (e.g. geopolitical or touristic documents) [7].

Our proposed framework assumes that the statistical dependencies of information pieces (and therefore their proximity) may also be determined by the user context. Future work will include an empirical verification of how our similarity function combines with appropriate representation spaces to result in operational similarity models. However, this issue is outside the scope of the present thesis.

Part III

Expirical Studies: Heterogenous Feature Aggregation and Ranking Fusion

6

Feature Aggregation in On-line Reputation Management on Twitter

6.1. Introduction

The explosion of available textual information on social media has opened the door to multiple applications, such as opinion mining, trend analysis or story detection. In this chapter, we take the online reputation monitoring (ORM) scenario as study case for the representation framework presented in this thesis. In this scenario, social media messages are analysed to identify conversations or events that can affect the reputation of a company or brand. In particular, this study case focuses on Twitter, which is probably the most common source for reputation monitoring. The following are two examples of tweets that mention the company BMW.

```
@Usedcarexpert:  BMW fires a new six-shooter http://torq.at/bf7
@yajnworb:  BMW for sale,anyone interested?
```

According to the dimensions defined by the Reputation Institute [99], these tweets can be classified as "products & sevices". In a finer granularity level, they can be associated with subtopics "new products" and "BMW vehicles for sale", respectively. In other words, we have a predefined finite set of dimensions (categories), for which we can generate a training dataset. On the other hand, the subtopics are dynamic. They evolve over time. Therefore, it is not enough to learn previously observed subtopics. The goal in our study case consists in grouping tweets according to subtopics. This is at the end a clustering problem and the core step is computing pairwise tweet similarity. That is, to what extent two tweets belong to the same subtopic or not.

In this scenario, the most traditional way of representing tweets is as a sequence of words. These are discrete and intrinsic tweet features. However, much more information is available from external systems. Having, for instance, a Bayesian dimension classifier, we have the membership value regarding each subtopic previously annotated in a training dataset. The problem is that the subtopics previously identified do not

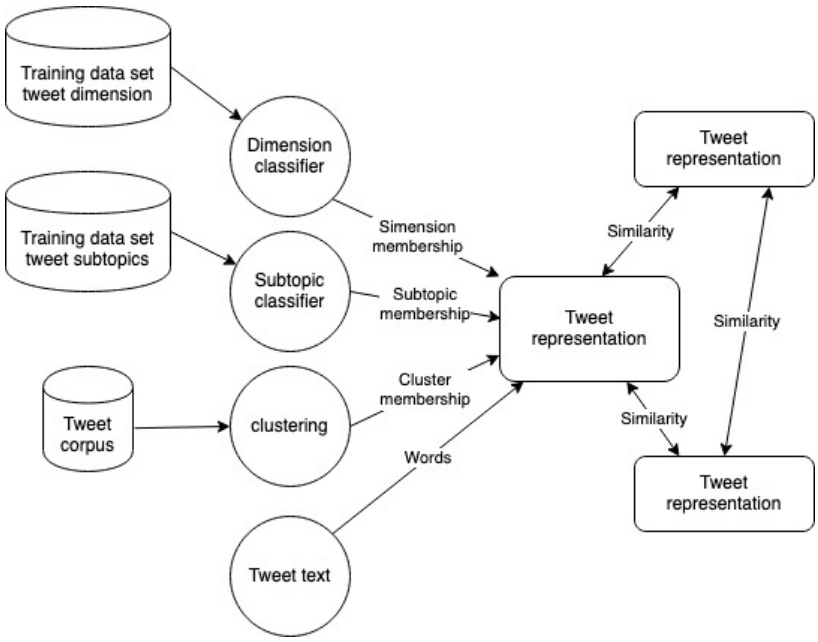

Figure 6.1

necessarily include emerging subtopics. However, it would be desirable to include this information in the tweet representation. In addition, we can extend this reasoning for other tweet categories for which training samples are available. For instance, we could consider the dimension membership value previously learned. Notice that the dimension is not the only feature to be compared when computing tweet similarity, but an additional characteristic. That is, we could combine intrinsic features (words) with extrinsic features generated by a supervised system to characterise the tweet. In addition, we could consider information from unsupervised learning. For instance, we could integrate in the tweet representation the proximity to tweet clusters generated previously by an external system.

Figure 6.1 illustrates this scenario. The working hypothesis is that integrating heterogeneous features (supervised vs unsupervised, intrinsic vs. extrinsic, etc.) can contribute to the information organization and similarity estimation in the context of ORM. Extrinsic features, such as supervised categorisation or proximity to clusters, capture dependencies and knowledge that is inferred from previous data. On the other hand, intrinsic features such as words capture new entities, events or terms that can affect the reputation of the company. The challenge is how to combine word features (binary) with quantitative signals generated by the classifier and the clustering system. In this chapter, we apply the Observational Representation Framework to overcome this challenge.

The experiments described in this chapter demonstrate that adding all these heterogenous features progressively increases the performance of tweet similarity computing. This result is verified under various similarity functions (Pointwise Mutual Information, Jaccard and Lin's distances and the Information Contrast Model proposed in this thesis). To the best of our knowledge, this is the first attempt to combine these kinds of signals in an unsupervised manner.

6.2. Dataset

We chose the RepLab dataset [6], which uses Twitter data in English and Spanish (more than 142000 tweets). The RepLab2013 corpus provides a collection of tweets that were obtained by searching for the names of various companies and entities (e.g., *BMW*, *Bank of America* and *Oxford University*). The corpus includes 61 entities and for each entity, at least 700 and 1500 tweets were collected for the training and test sets, respectively. The corpus also contains additional background tweets for each entity (up to 50000 tweets).

The tweet topic in the training and test datasets were manually annotated. Some of these topics are organizational (e.g., *"customer feedback"*) while others correspond with a particular event (e.g., *"Bank of America Chicago Marathon"*). The topics in the training dataset do not necessarily correspond to the topics in the test dataset. Therefore, applying classification techniques in isolation is insufficient. Furthermore, the training topics contain, in most cases, just a few tweets.

Additionally, each tweet was manually categorized with respect to standard reputation dimensions [99], which were the same for all the entities (e.g., *performance*, *leadership*, *innovation*).

6.3. Similarity Functions

The objective is to compare several representation models when estimating tweet closeness. OIQ converges to various representation models depending on the statistical assumptions about features. In this experiment we compare six representation approaches (corresponding with each assumption) and study the effect of adding new features (supervised and unsupervised). In particular, we study the effect of adding as features the proximity to previously annotated topics, categories (supervised features), and clusters (unsupervised features).

We aim to evaluate the relative effectiveness of representation models regardless of the similarity measure that is used. In these experiments, we consider three similarity schemes:

$S_{JACCARD}$: It computes the ratio between the intersection and the union of OIQs.

$$S_{JACCARD} = \frac{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1) \cap \mathcal{O}_\Gamma(d_2)\big)}{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1) \cup \mathcal{O}_\Gamma(d_2)\big)} \ .$$

Considering words as features and assuming independence and equiprobability, $S_{JACCARD}$ is equivalent to the original Jaccard similarity measure.

S_{LIN} : The second scheme considers the sum of OIQs instead of the union.

$$S_{LIN} = \frac{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1) \cap \mathcal{O}_\Gamma(d_2)\big)}{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1)\big) + \mathcal{I}\big(\mathcal{O}_\Gamma(d_2)\big)} \ .$$

We will refer to this as scheme S_{LIN}. According to Proposition 3.6, when considering words as features and assuming independence, it is equivalent to the original Lin's distance (see Subsection 2.3).

S_{ICM_β} : Finally, the third similarity measure is the information contrast model (ICM: see Subsection 2.6):

$$S_{ICM_\beta} = \mathcal{I}\big(\mathcal{O}_\Gamma(d_1)\big) + \mathcal{I}\big(\mathcal{O}_\Gamma(d_2)\big) - \beta \cdot \mathcal{I}\big(\mathcal{O}_\Gamma(d_1) \cup \mathcal{O}_\Gamma(d_2)\big) \ .$$

Fixing the parameter $\beta = 1$ and considering discrete features as words, it is equivalent to the traditional PMI. We experiment with the values $\beta = 1$ ($S_{ICM_1} = S_{PMI}$) and $\beta = 1.5$ ($S_{ICM_{1.5}}$).

For computational reasons, for all similarity measures, we assume independence across features and information additivity, as expressed in Equation (3.2). Let $d = (w_1, \ldots, w_m)$

$$\mathcal{I}\big(\mathcal{O}_\Gamma(d)\big) = \sum_{\gamma \in \Gamma} \sum_{i=1}^{m} \mathcal{I}\big(\mathcal{O}_\gamma(w_i)\big) \ . \tag{6.1}$$

6.4. Representation Schemes

We have observed in the previous sections that the Observational Information Quantity (OIQ) converges to diferent representation schemes under different assumptions. In addition, it enables quantitative signals that are produced by learned (supervised and unsupervised) features to be captured. Each representation approach that is compared in our experiments corresponds to a particular way of computing $\mathcal{I}\big(\mathcal{O}_\gamma(w_i)\big)$ in Equation (6.1):

$\mathcal{I}\big(\mathcal{O}_{\Gamma_{tf}}(d)\big)$: First, the simplest approach considers occurrence-based word features and assumes equiprobability among words. It corresponds to the tf representation model according to Proposition 3.2. Let Γ_{tf} be the set of occurrence-based word features:

$$\mathcal{I}\big(\mathcal{O}_{\Gamma_{tf}}(d)\big) = \sum_{w \in \Gamma_{tf}} tf(w, d) \ .$$

$\mathcal{I}\big(\mathcal{O}_{\Gamma_{tf.idf}}(w)\big)$: The second representation approach considers occurrence-based word features, but does not assume equiprobability among words. It corresponds to the

tf.idf representation model according to Proposition 3.4. Let $\Gamma_{tf.idf}$ be the set of occurrence-based word features:

$$\mathcal{I}\big(\mathcal{O}_{\Gamma_{tf.idf}}(d)\big) = \sum_{w \in \Gamma_{tf.idf}} tf.idf(w,d) \quad .$$

$\mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{tf.idf}}(d)\big)$: In the third approach, we use essentially the same approach as $\mathcal{I}\big(\mathcal{O}_{\Gamma_{tf.idf}}(d)\big)$; however, to avoid noise, we apply different thresholds to the IQ. That is, for a fixed threshold, we truncate to zero the salience of a coordinate if it is under the threshold, as expressed in Equation (3.1). According to trial experiments, we set the threshold at 5.0 for the word features.

$$\mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{tf.idf}}(d)\big) = \sum_{w \in \Gamma_{tf.idf}} th(w) \cdot tf.idf(w,d) \quad .$$

where

$$th(w) \begin{cases} 1, & \text{if } -log(P(w)) \geq th \\ 0, & \text{otherwise} \end{cases} \quad .$$

$\mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{\mathbf{Top}}}(d)\big)$: An important advantage of our representation model is that we can integrate new features into the word-based feature set. In the fourth approach, we consider the information of a set of topics, which is denoted as Top. Tweets in the training dataset are categorized into topics. For this, we assume that each topic, $t \in$ Top, is a feature. We also assume information additivity across words and the feature value for each word is estimated as a conditional probability, which is denoted as $P(t \mid w)$, in a similar way to the Naive Bayes classification method. Specifically, given a topic, $t \in$ Top, and the vocabulary, $\mathcal{V}$, the projection of a word, $w \in \mathcal{V}$, for each annotated topic in the training corpus (topic feature):

$$t(w) = P\big(t \mid w\big) = \frac{P\big(w \mid t\big) \cdot P\big(t\big)}{P\big(w\big)} \simeq \frac{\frac{freq_t(w)}{|\text{Top}|} \cdot \frac{|\text{Top}|}{|\mathcal{C}|}}{\frac{freq(w)}{|\mathcal{C}|}} = \frac{freq_t(w)}{freq(w)} \quad ,$$

where $freq_t(w)$ stands for the number of occurrences of a word w in topic t, (that is, in the set of tweets of the training dataset that belong to the topic) and $freq(w)$ stands for the number of occurrences of a word in the corpus $\mathcal{C}$. After that, and according to Definition 3.4, the OIQ of a word, $w \in d$, given a topic feature, $t \in$ Top, is computed as follows:

$$\mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{\text{Top}}}(d)\big) = \sum_{t \in \Gamma_{\text{Top}}} \sum_{w \in d} -log\Big(P_{w' \in \mathcal{C}}\big(t(w') \geq t(w)\big)\Big) \quad .$$

In a frequentist manner, we compute the cumulative distribution of topic feature values across word occurrences in the whole background tweet corpus (up to 50000 tweets).

$\mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{\mathbf{Dim}}}(d)\big)$: Furthermore, OIQ enables us to integrate information from various sources. According to this, in addition to the topic training data, we consider the classification of texts into reputational dimension categories (e.g., *performance*, *leadership*, and *innovation*). We denote the set of dimensions by Dim. It is assumed that belonging to the same dimensional category constitutes additional evidence of similarity. For this, we also consider the projection of words into reputational dimensions that are annotated in the training corpus in the same way as in the case of topic training data.

$\mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{\mathbf{Clus}}}(d)\big)$: To also integrate unsupervised learned features, the sixth representation approach consists of applying the same technique over automatically generated clusters instead of topic or category training datasets. We denote the set of clusters as Clus. Each feature is given by the proximity to a cluster, $c \in$ Clus, which is estimated in a similar way to previous approaches. For this, we use the clustering output from the best system that took part in the RepLab 2014 competition. For all the topic-, dimension- and clustering-based features, we have used the OIQ threshold of 11.0, which systematically outperforms other thresholds over the three types of features.

6.5. Evaluation Benchmark

The similarity estimation task consists of predicting whether two tweets refer to the same topic. For each entity in the test dataset, we generate a random sample of 1000 tweet pairs. Among them, there are 500 tweet pairs (d, d') whose elements belong to the same topic (not necessarily the same topic for all tweet pairs), and the remaining 500 pairs are those whose elements belong to different topics. For each evaluated approach, we sort the 1000 similarity instances according to the similarity estimation. Finally, considering different ranking lengths (varying the parameter k from 1 to 1000), we calculate the *precision at k*, which is the ratio of tweet pairs that belong to the same topic (according to the gold-standard) within the k instances with the highest similarity according to the approach.

Some similarity measures can produce many ties. In that cases, the evaluation results could be biased by the arbitrary similarity instance sorting criterion. In order to avoid this bias, we define Precision at k in probabilistic terms and pairs with equal similarity in the measure count a half in the probability estimation. Being X the set of 1000 similarity samples, X_{rel} and X_{unrel} the sets of 500 related and non related tweet pairs respectively, and being f a similarity function $f : X \longrightarrow \mathbb{R}$:

$$Precision_k(f) = P\left(x \in X_{rel} \mid f(x) > th_k^f\right) .$$

where th_k^f is the similarity measure value such that the area of cumulative density

Table 6.1: Precision at 200, 500 and 800 for all the evaluated approaches. Rows represent feature configurations and columns the corresponding precisions.

Sim. scheme	Representation model	Precision at k **200.0**	**500.0**	**800.0**
$S_{JACCARD}$	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf}})$	0.863	0.652	0.546
	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.909	0.678	0.550
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.917	0.682	0.551
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}})$	0.927	0.691	0.553
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}})$	0.926	0.691	0.553
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Clus}}})$	**0.931**	**0.704**	**0.559**
S_{LIN}	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf}})$	0.862	0.652	0.546
	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.909	0.678	0.550
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.917	0.682	0.551
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}})$	0.927	0.691	0.553
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}})$	0.926	0.691	0.553
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Clus}}})$	**0.931**	**0.704**	**0.559**
$S_{ICM_{1.5}}$	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf}}(d))$	0.848	0.651	0.542
	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.910	0.674	0.548
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.917	0.679	0.550
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}})$	0.927	0.688	0.553
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}})$	0.925	0.689	0.554
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Clus}}})$	**0.928**	**0.698**	**0.556**
S_{PMI}	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf}})$	0.795	0.573	0.505
	$\mathcal{I}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.891	0.661	0.535
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}})$	0.896	0.665	0.531
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}})$	0.910	0.673	0.536
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}})$	0.908	0.673	0.535
	$\mathcal{I}^{th}(\mathcal{O}_{\Gamma_{tf.idf}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Top}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Dim}}}) + \mathcal{I}^{th}(\mathcal{O}_{\Gamma_{\text{Clus}}})$	**0.925**	**0.681**	**0.547**

distribution, from the maximum $f(x)$ to th_k^f, is $\frac{k}{|X|}$:

$$P(f(x) > th_k^f) = \frac{k}{|X|} \ .$$

Under this probabilistic definition, being x_k the k^{th} instance sorted by f (with any arbitrary criterion for ties), precision at k can be estimated as:

$$Precision_k(f) = \frac{1}{k} \cdot \left(\left| \{x \in X_{rel} : f(x) > f(x_k)\} \right| + \frac{1}{2} \left| \{x \in X_{rel} : f(x) = f(x_k)\} \right| \right) \ .$$

6.6. Results

Table 6.1 shows the results. Similarity measures are evaluated according to precisions at 200 (the accuracy of the top 20% of similarity instances), 500 and 800. Each row represents a combination of a representation approach and a similarity scheme. Each column represents a precision level.Our purpose is to evaluate the capability of the model to add more features without adding noise. In particular we are interested on the differences among representation models for every similarity scheme ($S_{JACCARD}$,

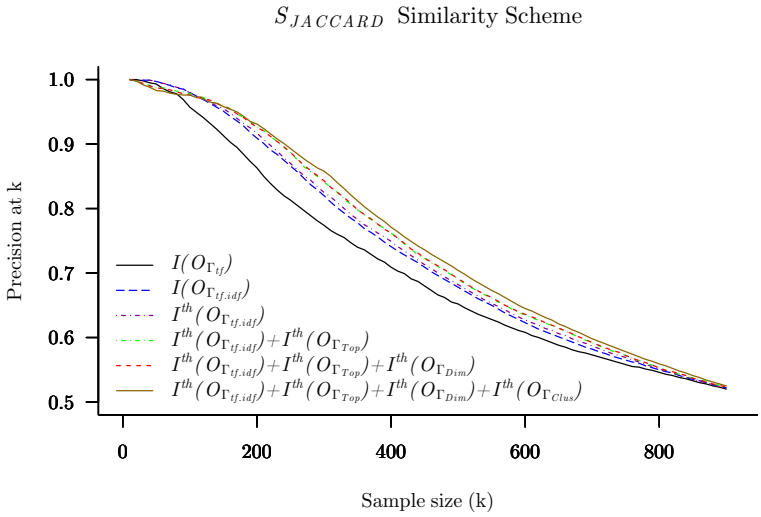

Figure 6.2: $S_{JACCARD}$ Similarity Scheme performance across different sample sizes (k). This scheme takes into account the union and the intersection of tweets.

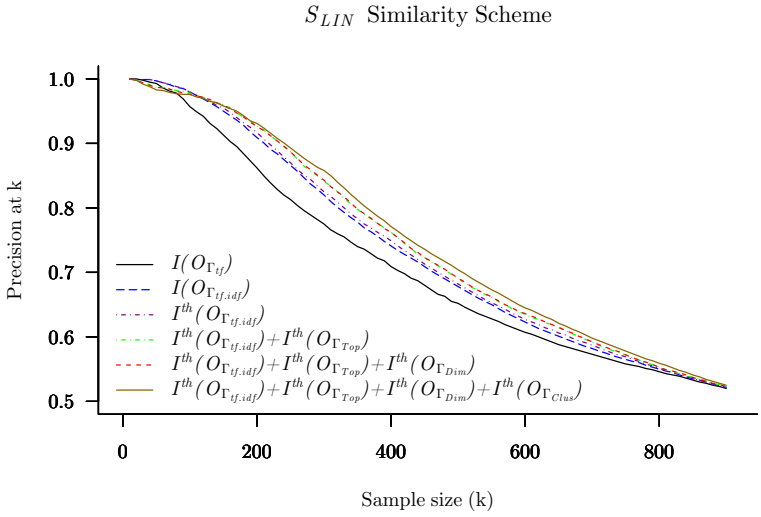

Figure 6.3: S_{LIN} Similarity Scheme performance across different sample sizes (k). Note the likeness with the $S_{JACCARD}$ Scheme due to their similar operation.

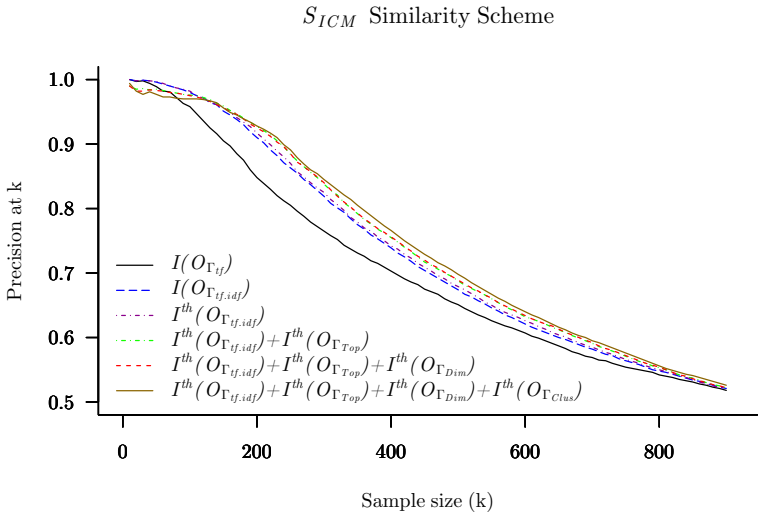

Figure 6.4: $S_{ICM_{1.5}}$ Similarity Scheme performance for different sample sizes (k). It shows the improvement when integrating heterogeneous continuous valued features.

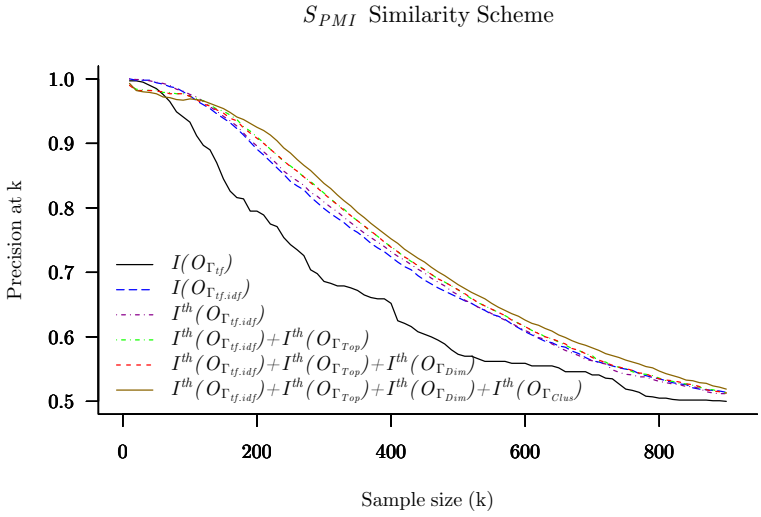

Figure 6.5: S_{PMI} Similarity Scheme performance for different sample sizes (k). It shows the consistency with the others Similarity Schemes.

S_{LIN}, $S_{ICM_{1.5}}$, and S_{PMI}), rather than the relative effectiveness of similarity measures. The results for models $S_{JACCARD}$ and S_{LIN} are identical. This is not an error. The differences appear after 4 decimals places. These two schemes are very similar.

The first observation is that $\mathcal{I}\big(\mathcal{O}_{\Gamma_{tf.idf}}(d)\big)$ outperforms $\mathcal{I}\big(\mathcal{O}_{\Gamma_{tf}}(d)\big)$ for all similarity schemes. There is an improvement of around 5% in precision at 200. That is, applying IQ and assuming different probabilities among words increases the effectiveness of the similarity estimation for every precision level and similarity scheme. This result is consistent with the findings of multiple works that corroborate the effectiveness weighting methods such as *tf.idf*. When adding a OIQ threshold, namely, $\mathcal{I}^{th=5}\big(\mathcal{O}_{\Gamma_{tf}}(d)\big)$, further improves the results. We can observe an improvement of around % in precision at 200. This is consistent with the effectiveness of stop-word removal and vocabulary reduction according to word frequency, which has been reported multiple times in the literature.

The addition of topic features (i.e. the proximity to topics in the training datasets) improves the performance around 5%, specially in precision at 500. The addition of dimension features (proximity to dimension tweet categories in the training dataset) does not produce any improvement. This suggest that the dimension of tweets is not necessarily related with their proximity in terms of dynamic topics. Finally, the addition of cluster features (proximity to clusters) improves slightly the similarity prediction performance.

Regardless the absolute quantitative improvements, the strength of these results is ground on: (i) they are consistent across similarity functions, (ii) adding non informative features do not affect the effectiveness substantially, (iii) the feature weighting does not require supervision and (iv) our theoretical framework provides a single formalization for diverse phenomena observed in the literature, such as the benefits of considering feature specificity (idf weighting), removing non informative features (stopwords), and (v) considering heterogeneous features improves similarity estimation performance, even when these features do not focus directly on the similarity evaluation target. In summary, the most relevant aspect of these experiments is the corroboration of the hypothesis studied in this chapter: **Adding heterogeneous features under the same OIQ-based weighting criterion increases progressively the similarity estimation performance, even when features include both discrete and continuous values and have different scale properties.**

As an additional experiment, Figure 6.2 shows the performance across k levels of precision with a fixed IQ threshold for the $S_{JACCARD}$ similarity scheme. Figures 6.3, 6.4 and 6.5 show similar results for the S_{LIN}, $S_{ICM_{1.5}}$ and S_{PMI} similarity schemes, respectively. These results suggest that the improvement when integrating heterogeneous continuous-valued features without supervision is consistent across similarity schemes and the k levels of precision. In addition, the performance increase due to additional features is especially visible at medium k levels. A possible reason is that classification methods do not contribute when there is a high similarity according to words. However, the indirect evidence is more effective at medium similarity levels.

6.7. Overcoming Limitations of Independence and Additivity

The model presented in previous sections has two main drawbacks: the need to assume *feature independence* and *information additivity*. One can view this representation model as an ensemble aggregating different scores (*independence*) in which aggregation is only performed for the sum operator (*information additivity*). To mitigate these effects, it seems that a tuning process is feasible. For instance, a potential solution could involve the use of an arbitrary function to combine features, which could assign weights in a supervised manner. The function can be expressed as a weighted sum:

$$w_1 \cdot \mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{tf.idf}}(d)\big) + w_2 \cdot \mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{\mathrm{Top}}}(d)\big) + w_3 \cdot \mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{\mathrm{Dim}}}(d)\big) + w_4 \cdot \mathcal{I}^{th}\big(\mathcal{O}_{\Gamma_{\mathrm{Clus}}}(d)\big) \ ,$$

where w_1, w_2, w_3 and w_4 could be learned from the training data. This question is also motivated by the fact that in any of the features analysed so far, the number of values they can take is arbitrary and defined by the user; for example, given a stream of tweets, the user is the one who defines the possible topics of the stream, and the more topics are defined, the more weight that feature will have on the sum. We will carry out two experiments to check this question:

1. Directly converting the ICM schema into a weighted sum of OIQs, i.e., parametrize the sum of OIQs of every feature and optimize the summation with several classification algorithms.

2. To study the effect of the weights directly on the initial model, i.e., to repeat the experiments of previous sections modifying single parameters. This way, we can observe the redundancy of one of the features, in particular, we have chosen the feature Topics, since the experiments are evaluated on it.

First, we conduct a brief statistical analysis of these datasets. We verify that the training dataset includes $9,088$ words. To determine the number of words used as features, we eliminate repeated words and are left with a total of $2,755$ words (30.3%). Of these words, 151 (5.5%) are stopwords. In the test dataset we have $20,066$ words, of which only $4,773$ (23.7%) are used as features, while 329 (6%) are stopwords. On the other hand, the test dataset (different from the training dataset) includes $3,628$ new words (76.01%), and the percentage represented by stopwords of these words is 42.29%. These data reflect dynamism found in the context of social media. With regard to topics, it must be noted that they are not categories, as in general topics of the training corpus are different from topics of the test corpus. The number of topics covered depends on each company even though the number topics averages at 58 in the training dataset, 50 in the test dataset and only 10 repeated in both datasets for all companies. This descriptive figures show the high dynamicity of the context; both words and topics are constantly changing.

6.7.1. Parametric Feature Weight Optimization

In the first experiment, we take advantage of the ICM scheme form defined in Section 6.3 to express its formula as a polynomial whose variables are the weights to be optimized.

$$S_{ICM} = w_1 \cdot S_{ICM}^{\Gamma_{tf.idf}} + w_2 \cdot S_{ICM}^{\Gamma_{\text{Top}}} + w_3 \cdot S_{ICM}^{\Gamma_{\text{Dim}}} + w_4 \cdot S_{ICM}^{\Gamma_{\text{Clus}}} \ ,$$

where each w_i represents the relative weight of each type of feature and $S_{ICM}^{\Gamma_{tf.idf}}$ stands for S_{ICM} using *tf.idf* as the only feature.

To perform this experiment, we represent each pair of tweets according to their similarities over each single feature. In particular, we consider words from the training dataset and use the similarities between both tweets considering each word as training features. As a result, we replace the Information Quantity of features defined in our model with a trained weighting criterion. In the calculation of similarities we use ICM since among the four similarity schemes presented in subsection 6.3, only ICM allows to decompose a similarity feature by feature, as it applies a formula whose operators are additions and subtractions. However, similarities of Jaccard and Lin do not allow for such decomposition due to their division operator. We use the same β values that we used in previous experiments.

Once we have calculated the similarities between each pair of tweets, we train the data using Naive Bayes, J48, SVM and Logistic Regression as classification algorithms. Then, we average the IQs for all companies. We select the dependent variable belonging to the same topic and the independent variables being the IQs of the features. The experiment is performed with Weka software [118]. The performance is shown in Table 6.2. As usual, the SVM classifier outperforms the other algorithms. However, regardless of the algorithm used, this success rate corresponds to the "Precision at 500" shown in Table 6.1 for the previous experiments, which achieve a higher degree of accuracy. We also carried out the same experiment while applying different filters to the data by, for example, eliminating those words with a frequency of less than four, but even more random classifiers were obtained.

Table 6.2: ICM schema as a weighted sum of IQs. It is shown the precision of different classification algorithms.

	Precission
Naive Bayes	56.1%
J48	55.8%
SVM	57.4%
Logistic Regression	53.8%

The results show that the polynomial coefficients are all practically equal to 1. We could infer that the relative effect of each feature is not so important, and this effect could be because the components of the sum themselves are based on Information Theory. Notice that the internal optimization function used by Weka (conditional log likelihood) is not exactly the one used in our evaluation framework. However, in both cases high similarity

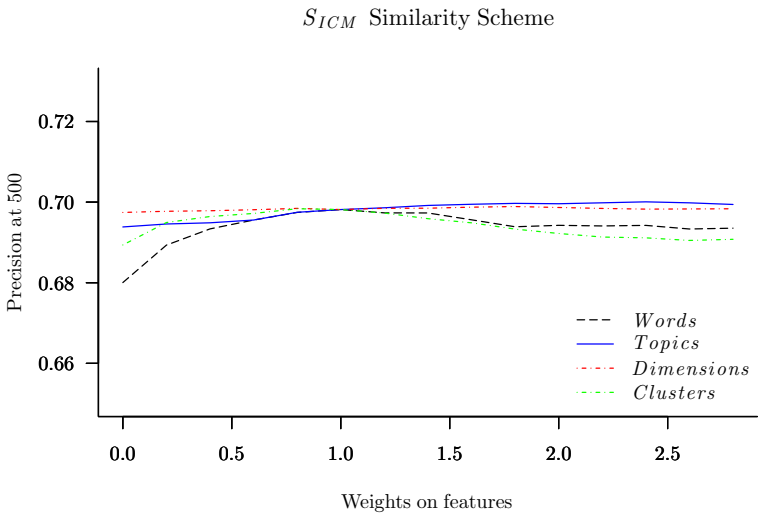

Figure 6.6: Representation of S_{ICM} Similarity Scheme performance across different weights assigned to features. On the x-axis it is possible to see the weight assigned to each feature, while keeping the value 1 for the rest of feature types, the reached precision at 500 can be seen on the y-axis.

values assigned to tweets pairs belonging to the same topic are rewarded. Another aspect to be taken into account is that we could also optimize weights over other (non linear) similarity schemes such as Jaccard or Lin. We let the empirical analysis of these aspects for future researches.

The precision achievement shown in Table 6.2 confirms the results of the RepLab evaluation campaign [6] in which the low performance of the classifiers is shown due to the high dynamicity of the context; both words and topics are constantly changing.

6.7.2. Robustness Analysis

In the second experiment, we study the influence of the weights directly on our evaluation framework. For each similarity scheme (ICM, Jaccard and Lin) we repeat the experiments of Subsection 6.6 by modifying the weight of each feature type. The results are shown in Figures 6.6, 6.7 and 6.8: on the x-axis, we have the value of the weights given to each feature, while keeping the value 1 for the rest of the feature types (ablation analysis), and on the y-axis, we can see the precisions at 500.

Figure 6.6 shows that increasing the weight of the features decreases accuracy in the cases of Words and Clusters. It means that these features contribute more to the performance than the other two features (this fact was not obvious from the results reported in the previous section). In the case of Dimensions, the precision remains somewhat stable, and only in the case of Topics there is a small improvement. This outcome occurs because we are evaluating Topics, and the feature is more closely related to the answer. This slight improvement is limited by the robustness of the model when all weights are equal to 1.

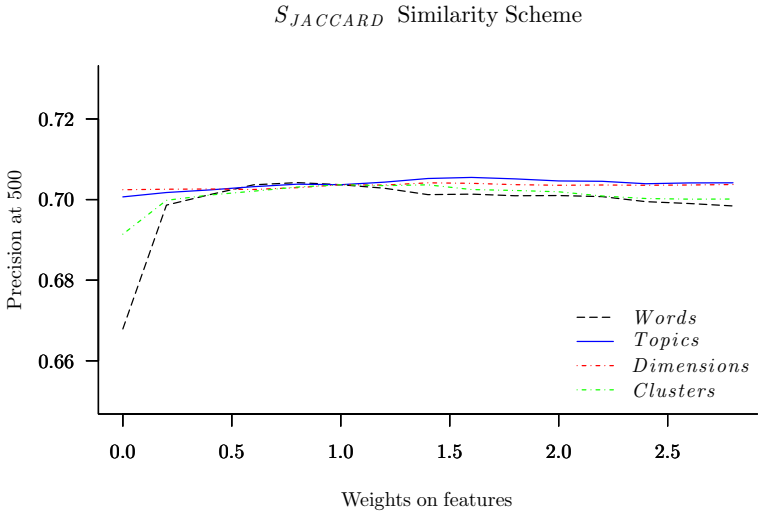

Figure 6.7: Representation of $S_{JACCARD}$ Similarity Scheme performance across different weights assigned to features. On the x-axis it is possible to see the weight assigned to each feature, while keeping the value 1 for the rest of feature types, the reached precision at 500 can be seen on the y-axis.

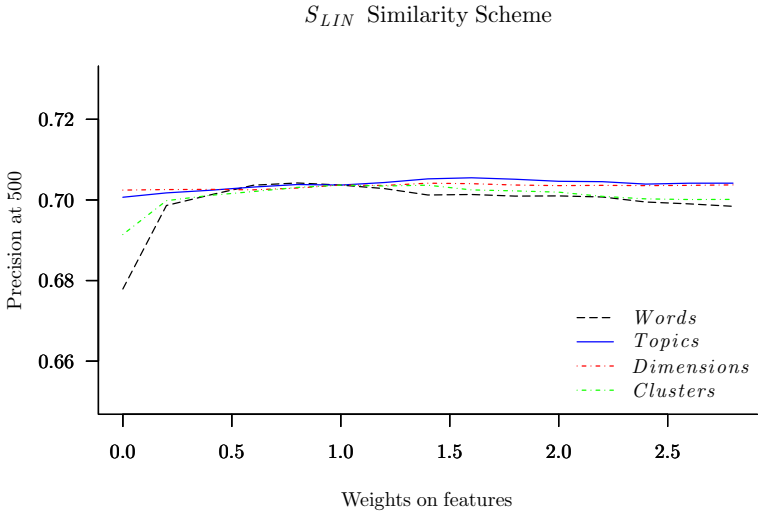

Figure 6.8: Representation of S_{LIN} Similarity Scheme performance across different weights assigned to features. On the x-axis it is possible to see the weight assigned to each feature, while keeping the value 1 for the rest of feature types, the reached precision at 500 can be seen on the y-axis.

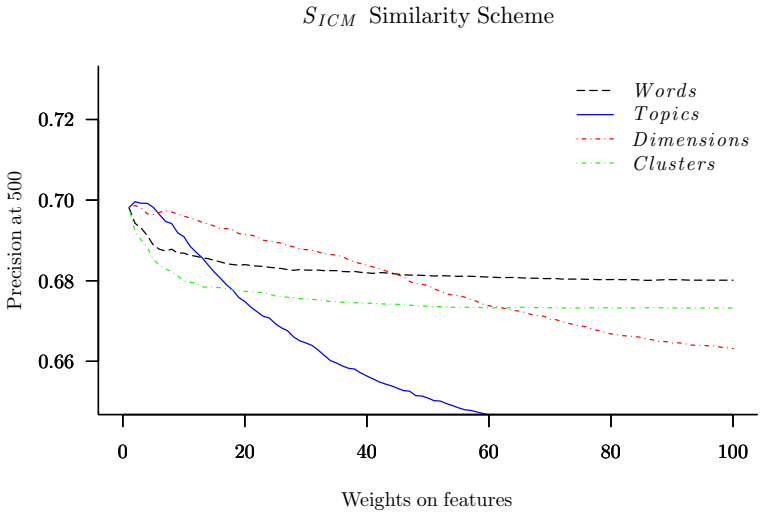

Figure 6.9: Representation of the robustness of the weighted model for the S_{ICM} Similarity Scheme. On the x-axis the weight of Topis is increased, while the precision at 500 is observed on the y-axis.

A further step would be to determine how much redundancy would have to be added to Topics so that the effect of the rest is annulled. Figures 6.9, 6.10 and 6.11 show the parameter of features on the x-axis for a wider range of values than the previous figures, and on the y-axis the precision at 500. If we add a weight of 100 to the feature Topics while keeping the rest of features equal to 1, we obtain an accuracy of 63% for the ICM similarity scheme, which is close to the precision considering Topics as the only feature (61%); i.e., great redundancy is needed to match the action of the feature Topics in isolation.

Independence and *information additivity* are two open issues to be solved in our model. These experiments show us the difficulties associated with dealing with this concern, and an interesting future task would be to tackle this issue by modelling dependence between features.

6.8. Conclusions: The Aggregation of Heterogeneous Features via ORF

In the case of study described in this chapter, we have estimated tweet similarity in order to infer fine grained subtopics in the reputation information stream. We are interested in integrating intrinsic features (i.e., words) with supervised and unsupervised extrinsic features (i.e. tweet categories and clusters) in order to exploit previous data while addressing the dynamic nature of information flows in reputation management. The proposed representation framework (ORF and OIQ) and the similarity function derived in Chapters 4 and 5 has been the base on which the experiments have benefited.

The experiments showed that, adding quantitative features, such as dimensions or prox-

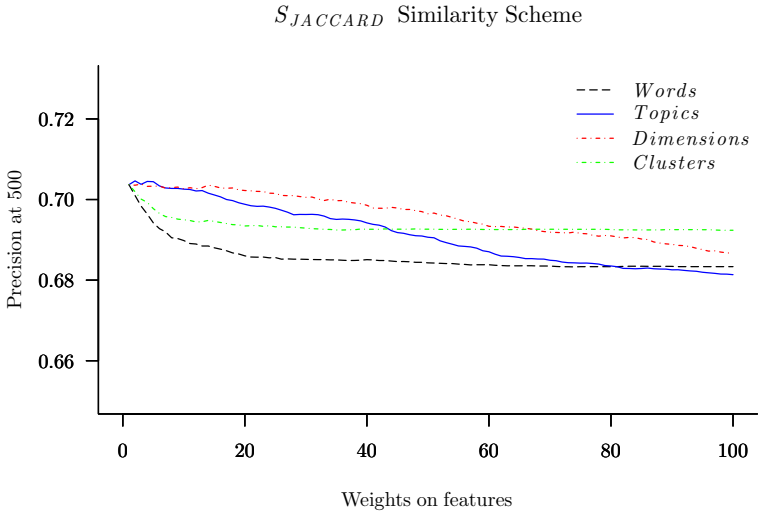

Figure 6.10: Representation of the robustness of the weighted model for the $S_{JACCARD}$ Similarity Scheme. On the x-axis the weight of Topis is increased, while the precision at 500 is observed on the y-axis.

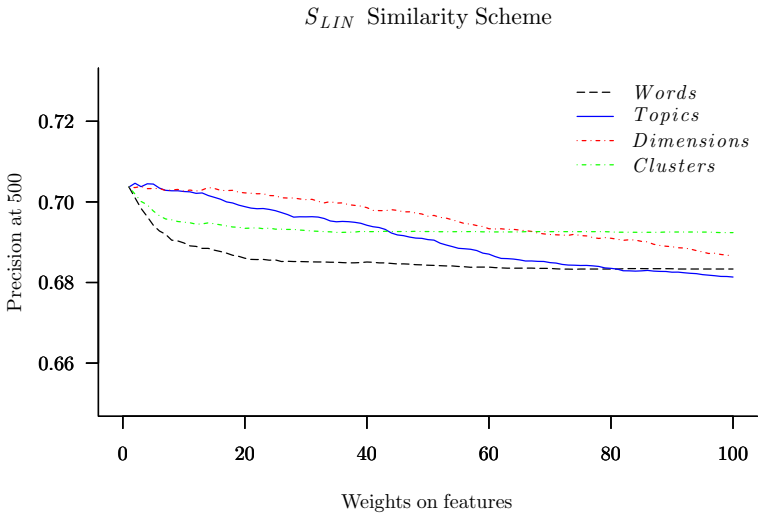

Figure 6.11: Representation of the robustness of the weighted model for the S_{LIN} Similarity Scheme. On the x-axis the weight of Topis is increased, while the precision at 500 is observed on the y-axis.

imity to clusters, under the OIQ framework without supervision increases the tweet similarity prediction effectiveness. Note that these features are not directly related to the similarity target, e.g., proximity to previously seen topics, reputational dimensions or automatically generated clusters. The effectiveness improvements are consistent across similarity functions and evaluation metrics. In addition, the experiments confirm that ORF provides a single theoretical basis and formalization for popular and effective representation techniques such as tf.idf weighting or stopword removing.

On this line, there are many aspects to be investigated in future works. Maybe the most inmediate issue is to prevent the feature independence assumption. At word feature level, this limitation has been tackled in recent years via dimensionality reduction techniques such as word or phrase embeddings. An open question derived from this study is whether we can combine these representation techniques with heterogeneous features under the ORF and OIQ framework.

7

Applying the Observational Information Framework to Ranking Fusion

7.1. Introduction

Ranking fusion (or aggregation) consists of combining multiple rankings into a single ranking. Under the assumption that different signals or information sources reflect different aspects of a ranking problem, ranking fusion attempts to optimize the performance of their combination. The ranking fusion problem is encountered in many research areas, such as averaging methods, meta-classifiers, feature analysis, information retrieval, and the combination of evaluation measures [116, 119, 110, 90, 109, 10, 15, 19, 70]. The unsupervised combination of rankings through voting is a powerful strategy. Empirical studies addressing various scenarios have repeatedly corroborated that voting methods offer a performance equivalent to the best individual measure. In addition, voting methods avoid overfitting and guarantee robust results across different data sets (compared to supervised strategies), even if the optimal measure is different for each data set.

In this chapter, we study the application of the Observational Representation Framework (see Chapter 3) to ranking fusion. In particular, we study whether, when item scores (rankings) are considered as features, the Observational Information Quantity (OIQ) behaves as a ranking fusion function and explains the strengths and weaknesses of existing approaches.

On the basis of previous work, we first define a set of desirable formal properties for ranking fusion. We then analyse whether the best existing ranking fusion methods satisfy them. We observe that when item scores are taken as features, the OIQ satisfies all these properties, and moreover, we identify under which conditions existing ranking fusion algorithms approximate the OIQ. We empirically study the behaviour of different OIQ estimation approaches over six ranking fusion benchmarks.

7.2. Desirable Properties

In this section, we formalize the ranking fusion problem and define a set of formal properties extracted from empirical practices in previous ranking fusion studies. Given a set of items, $\mathcal{D}$, a ranking, denoted by r_γ, is a list of items in $\mathcal{D}$ sorted in accordance with a certain score function, $\gamma : \mathcal{D} \longrightarrow \mathbb{R}$. We will use $rank_r(d)$ to denote the ranking position of d in the ranking r. Rankings can be categorized into three types [10, 51, 40]: total, partial, and top-k rankings. Here, we focus on total ranking, in which one item is assigned to each ranking position ($\gamma(d) \neq \gamma(d'), \forall\, d, d' \in \mathcal{D}$). Consider a set of available rankings generated on the basis of different information sources (or signals) in Γ. The ranking fusion function denoted by $F_\Gamma : \mathcal{D} \longrightarrow \mathbb{R}$ returns a score for each item on the basis of its ranking positions.

There is a common intuition that a high ranking position represents positive evidence regardless of the other rankings with which it is to be combined, i.e., the more high ranking positions an item achieves, the higher a position it should receive in ranking fusion. This can be stated as a monotonicity property.

Property 7.1 (Strict Monotonicity). *Moving an item up in one ranking increases its score as assessed by the fusion function. When $\gamma_{d_1 \leftrightarrow d_2}$ the result of swapping the scores of d_1 and d_2 in γ is verified as follows:*

$$\gamma(d_1) > \gamma(d_2) \Rightarrow F_{\Gamma \cup \{\gamma\}}(d_1) > F_{\Gamma \cup \{\gamma_{d_1 \leftrightarrow d_2}\}}(d_1)\ .$$

Note that, when item scores are taken as feature values, this property is similar to Property 3.1, which states that increasing the feature instantiation values increases the OIQ.

The second issue is the treatment of redundant signals. When systems that have been artificially generated are combined, equal signals do not provide additional information and should be considered redundant. This is not the case for voting systems, since the majority vote determines the final combined outcome method; however, this property must be taken into account if we wish to consider minority votes. For instance, in the context of meta-classifiers, multiple authors have shown that combining heterogeneous classifiers increases the accuracy of the combined result [94, 70]. In regard to ranking fusion, Vogt et al. found that the effectiveness increases when relevant documents are ranked in a different fashion [116]. More recently, it has been shown explicitly that the diversity of the system outputs is related to the diversity of the documents in the fused rankings [119]. In the prediction of IR system effectiveness[1], Spoerri found that selecting systems from different research teams when generating *pseudo-qrels* increases the predictive power [110]. Nuray et al. found that selecting systems that differ from the norm is also beneficial [90]. In the context of combining machine translation evaluation measures (representing the similarity between system outputs and human-generated references), the benefits of combining heterogeneous similarity-based measures have been

[1] IR system effectiveness prediction consists on replacing manual relevance annotation with a set of highly ranked documents for evaluation purposes (*Pseudo-QRels*).

observed on multiple occasions [5, 79, 4, 25]. More specifically, it has been observed that mixing text similarity measures based on diverse linguistic levels (lexical, syntactic) usually improves the results. This is stated by the next property, which is related to Property 3.5 concerning the OIQ.

Property 7.2 (Dependence). *Redundant rankings do not affect the fusion function. Given two signals, γ_1, γ_2, if there exists a real strict monotonic function g that satisfies $\gamma_1(d) = g\big(\gamma_2(d)\big),\ \forall\ d \in \mathcal{D}$, then, $F_{\{\gamma_1\}}(d) = F_{\{\gamma_1,\gamma_2\}}(d)\ \ \forall\ d \in \mathcal{D}$.*

Finally, when combining a pair of contradictory rankings, i.e., rankings ordered with the items in opposite positions to each other, the ranking fusion function does not have any criteria for determining preference between items. Therefore, the fusion function should be constant across items.

Property 7.3 (Cancellation). *The fusion function for two inversely valued rankings is a constant function. Given two inverse signals γ_1 and γ_2, if for every pair of items $d,\ d' \in \mathcal{D}$, we have:*

$$\gamma_1(d) > \gamma_1(d') \Longleftrightarrow \gamma_2(d) < \gamma_2(d')\ ,$$

then $F_{\{\gamma_1,\gamma_2\}}(d) = k$.

Note that this property has some similarities with Property 3.7 concerning the OIQ.

7.3. Ranking Fusion Algorithms

Various methods exist for merging rank-ordered lists [15, 39, 72, 120]. Basically, they use information that is readily available from ranked lists of items, i.e., the ordinal rank assigned to each item in each ranking list or some transformation of those ranks. The simplest way to achieve unsupervised fusion consists of applying traditional *averaging schemes*, such as the arithmetic and geometric means, or the maximum or minimum.

$$Avg_\Gamma(d) \propto \sum_{\gamma\in\Gamma}\big(\gamma(d)\big) \qquad Geo_\Gamma(d) \propto \left(\prod_{\gamma\in\Gamma}(\gamma(d))\right)^{\frac{1}{|\Gamma|}}$$

$$Max_\Gamma(d) \propto \max_{\gamma\in\Gamma}\big(\gamma(d)\big) \qquad Min_\Gamma(d) \propto \min_{\gamma\in\Gamma}\big(\gamma(d)\big)$$

In the case of the arithmetic and geometric means, a change in any single ranking produces a change in the combined result. Therefore, they exhibit *monotonocity*[2]. This is not the case for the maximum and minimum schemes, which obviously are not strictly monotonic with respect to the maximum or minimum measure score. Only the arithmetic mean is strictly monotonic [3]. Moreover, averaging schemes do not satisfy the

[2] In the case of geometric mean, we need to assume that $\gamma(d) > 0,\ \forall d \in \mathcal{D}$.

[3] Note that a zero value avoids the effect of the rest of rankings in geometric and harmonic means, and maximum and minimum only consider at the end one of the combined features.

dependence property. If a ranking is replicated indefinitely, the combined score calculated with in the arithmetic, geometric or harmonic mean will tend to more strongly resemble the replicated score. However, this is not the case for the maximum and minimum functions. These combination schemes are not sensitive to redundant features. These score-based approaches also do not comply with *cancellation*. The reason is that this property is defined over opposite ordinal rankings, whereas these fusion methods rely on the original scores.

In the context of information retrieval, the most popular baseline for document ranking fusion is perhaps the Reciprocal Rank Fusion (RRF) method proposed by Cormack [24]. This method considers the inverse of the item positions in a ranking, which is adjusted by a specific constant. Thus, this method that relies on the ranking positions instead of the original scores. The Reciprocal Rank Fusion score is defined as follows:

$$RRF_\Gamma(d) = \sum_{\gamma \in \Gamma} \frac{1}{rank_\gamma(d) + k} \ .$$

RRF is strictly monotonic and sensitive to changes in individual rankings. However, it does not comply with *dependence*, as it is also sensitive to redundant rankings. RRF also does not satisfy *cancellation*. The reason is that RRF manages each individual ranking independently, and documents at the top in each ranking are especially influential. Formally, it is sufficient to see that the expression $\frac{1}{rank_\gamma(d)+k} + \frac{1}{|\mathcal{D}|-rank_\gamma(d)+k}$ is not constant.

Another closely related research area is the study of voting rules based on preferences. In fact, a ranking fusion function can be seen as a preference function that takes the voter preferences as input. In binary voting, that is, in a case with two alternatives (items), the use of the simple majority voting rule has always been unquestionable [80]. However, when preferences for multiple items must be considered, formulating a voting rule is not straightforward. The core of the problem is the impossibility theorem [121], which establishes that majority voting is not necessarily transitive. This means that one item can be considered an improvement over another when they are compared directly in terms of majority voting, but the opposite result may be obtained in a corresponding merged ranking. Such cycles of majority preference can be solved by different preference systems.

The most popular voting system for preferences is the Borda count algorithm [31]. Borda's rule is a voting procedure in which the agents must order the different candidates linearly, in accordance with their merit in each case, and assign them correspondingly graduated scores to determine the highest total count obtained, which indicates the winning candidate. Let us define $r_\gamma(d)$ as the salience of document d in the ranking produced by γ:

$$r_\gamma(d) = \left| \left\{ d' : rank_\gamma(d) \geq rank_\gamma(d') \right\} \right| \ .$$

The classical individual Borda count for an item d and a ranking r_γ is given by $r_\gamma(d)$;

then, the total or aggregated Borda count for a set of rankings is given by:

$$Borda_{\Gamma}(d) = \sum_{\gamma \in \Gamma} r_{\gamma}(d) \ .$$

This approach is equivalent to the arithmetic average of the rankings when the rankings are transformed into ordinal scales. Thus, the Borda count possesses the same properties as the average rank. It complies with *monotonicity*. For instance, if one item, d, improves over another item, d', in all rankings (and strictly for some of them), then the number of items improved upon by d is at least the same as that for d' plus one. Borda's rule assumes that each voter has an intrinsic value, even if it is redundant; thus, it fails to satisfy *dependence*. Regarding *cancellation*, Borda's rule satisfies this property since it yields the same result for every pair of items.

As in the case of scor-based fusion methods, we can also consider the minimum rank and maximum rank schemes.

$$MaxRank_{\Gamma}(d) \propto \max_{\gamma \in \Gamma} \left(r_{\gamma}(d) \right) \qquad MinRank_{\Gamma}(d) \propto \min_{\gamma \in \Gamma} \left(r_{\gamma}(d) \right)$$

Again, these schemes satisfy *dependence* but not *monotonicity* or *cancellation*.

Another family of ranking fusion rules is *majoritarian rules*, in which preferences are expressed between pairs of alternatives. These preferences are aggregated by means of concrete majority rules yielding a peer-to-peer comparisons of the alternatives, as in the family of voting methods called *Condorcet methods* [32, 87]. Condorcet noticed that in an election carried out via the simple majority method, individuals can rationally manifest their preferences, but this rationality may be lost in the aggregate (this is known as the paradox of the vote, the Condorcet effect or the existence of cycles).

The pure Condorcet voting method avoids the Condorcet paradox by identifying items belonging to the same transitivity cycle of majority voting and ranking them together. However, it is very costly in computational terms because it requires detecting all cycles. The Copeland Condorcet variant considers how many items are improved by considering a majority of rankings, instead of avoiding differences in cycles as in Condorcet's rule. The score value of an item d in the Condorcet method can be written as:

$$Copp_{\Gamma}(d) \propto \left| \left\{ d' : \left| \{\gamma : r_{\gamma}(d) \geq r_{\gamma}(d')\} \right| > \left| \{\gamma : r_{\gamma}(d') \geq r_{\gamma}(d)\} \right| \right\} \right| \ .$$

This family of voting methods includes several alternative approaches [32, 87], such as the ranked pairs, Copeland, Kemeny−Young, minimax and Schulze approaches.

In general, the Condorcet method and its variants do not strictly satisfy monotonicity, as an increase in ranking position does not necessarily produce a change in the outcome of pairwise majority voting. The Condorcet method assumes that each voter has an intrinsic value; thus, it does not satisfy *dependence*. However, the Condorcet method does *cancellation* since there is a voting tie for every pair of items.

Table 7.1: Formal Comparison of Unsupervised Combining Signals.

Method	**Monotonocity**	**Dependence**	**Cancellation**
Arithmetic and geometric means	✓	–	–
Maximum and Mininum	–	✓	–
Reciprocal Rank Fusion	✓	–	–
Borda Count	✓	–	✓
Maximum and minimum rank	–	✓	–
Condorcet	–	–	✓
UIR	–	✓	✓

The Unanimous Improvement Ratio (UIR) [9] was originally applied to compare two systems s_1 and s_2 in terms of a set of metrics Γ given a set of test cases C. Basically, the UIR counts in how many cases the first system outperforms the second in terms of every metric simultaneously:

$$UIR(s_1, s_2) = \frac{\left|\{c : (\gamma(s_1) \geq \gamma(s_2)),\ \forall\ \gamma\}\right| - \left|\{c : \gamma(s_1) \leq \gamma(s_2)),\ \forall\ \gamma\}\right|}{|C|} \ .$$

It is possible to adapt the UIR to the ranking fusion scenario in the following way[4]. Given an item d, with all other items considered as test cases, the metrics are the score functions $\gamma \in \Gamma$ that generate the item rankings:

$$UIR(d) = \frac{\left|\{d' \in D : \gamma(d) \geq \gamma(d'),\ \forall\ \gamma\}\right| - \left|\{d : \gamma(d) \leq \gamma(d'),\ \forall\ \gamma\}\right|}{|\mathcal{D}|} \ .$$

$UIR(d)$ approximates the following probability:

$$P\big(\{d' \in \mathcal{D} : \gamma(d) \geq \gamma(d'),\ \forall\ \gamma\}\big) - P\big(\{d' \in \mathcal{D} : \gamma(d) \leq \gamma(d'),\ \forall\ \gamma\}\big) \ .$$

In essence, the UIR is similar to the Condorcet method, but the pairwise comparisons are based on unanimous improvement rather than majority voting. In fact, the Copeland approach and the UIR converge when only two rankings are considered. However, the UIR is less sensitive to monotonic differences than the Condorcet method. The reason is that outperforming another item in every ranking simultaneously is a very restrictive condition. Therefore, it is very possible to move up in all rankings without affecting the UIR.

On the other hand, the main strength of the UIR is that it satisfies *dependence*. Because it is based on unanimous rank improvements, adding redundant rankings does not affect the combined result. The UIR also satisfies *cancellation*, since it yields the same result for each pair of oppositely ranked items. The desirable properties verified by each

[4]It is grounded on the fact that an unanimous improvement is the only Conjoint Relational Structure which is independent from any weighting parameter.

method of ranking fusion are summarised in Table 7.1.

7.4. The OIQ as a Ranking Fusion Algorithm

The OIQ behaves as a ranking fusion algorithm when items are considered as documents and the scores function as the features, Γ. That is, the projection $\pi_d(\gamma)$ of feature γ onto document d is $\gamma(d)$. We also need to consider that the ranked items in the set $\mathcal{D}$ are random samples of the universe of items $\mathcal{D}^u$ whose feature values follow a certain distribution. Then, the OIQ-based ranking fusion algorithm is expressed as follows:

$$\mathcal{I}_\Gamma(d) = log\left(\frac{1}{P\big(\{d' \in \mathcal{D}^u : \gamma(d') \geq \gamma(d), \forall\ \gamma \in \Gamma\}\big)}\right) .$$

In other words, given a document $d \in \mathcal{D}$, the OIQ considers the probability that documents in $\mathcal{D}^u$ will surpass d in every ranking.

An important contribution of the OIQ relative to previous ranking fusion approaches is that, whenever the score distribution in D^u is known, it satisfies the three desirable properties described in Section 7.2.

Proposition 7.1. *The OIQ based ranking fusion approach satisfies monotonicity, dependence and cancellation.*

Therefore, according to our analysis of the literature, the OIQ is the only ranking fusion approach that satisfies these three constraints. However, computing the OIQ requires knowing the continuous distribution of the features across items. If this distribution is unknown, it is necessary to estimate it from the available items in the rankings. Under the assumption that the items in $\mathcal{D}$ are a randomly sampled set from $\mathcal{D}^u$, the OIQ can be formulated as follows:

$$\mathcal{I}_\Gamma(d) = log\left(\frac{N}{\big|\{d' \in \mathcal{D} : rank_\gamma(d') \leq rank_\gamma(d), \forall\ \gamma \in \Gamma\}\big|}\right) ,$$

where N represent the ranking length. This estimation is closely related to the UIR. The correspondence can be seen in Proof A.4.2

Proposition 7.2. *The UIR has the following correspondence with the OIQ:*

$$UIR(d) \simeq 2^{I_{\{\gamma_1,\ldots,\gamma_n\}}(d)} - 2^{I_{\{-\gamma_1,\ldots,-\gamma_n\}}(d)} .$$

As in the case of the UIR, if the collection of documents is insufficient, i.e., $|\mathcal{D}| < |\Gamma|$, then the OIQ will not correctly capture the document relations and *monotonicity* will not be satisfied. The reason is that there may not be enough document to observe a new unanimous feature score improvements when an item moves up a ranking position. However, the *dependency* and *cancellation* properties are preserved.

If we assume feature independence, then the OIQ can be estimated as follows:

$$\mathcal{I}_\Gamma(d) = \sum_{\gamma\in\Gamma} log\left(\frac{N}{P\big(\{d'\in\mathcal{D} : rank_\gamma(d')\leq rank_\gamma(d)\}\big)}\right) .$$

This OIQ approach can be expressed as:

$$\mathcal{I}_\Gamma(d) \propto -\sum_{\gamma\in\Gamma} \log\left(N - r_\gamma(d)\right) .$$

Therefore, this OIQ estimation shows a correspondence with Borda's rule. Essentially, it consists of averaging the ranking positions but on a logarithmic scale. Similar to the Borda count, when independence is assumed, this OIQ formulation satisfies *monotonicity* at the cost of *dependence* and *cancellation*.

Copeland's method is based on the probability that the number of features that overcome an item is higher than the number of features that do not overcome this item. The Copeland method can be interpreted as a smoothed version of the OIQ in which unanimous improvement is replaced by majority voting:

$$\begin{aligned} Copp_{\mathcal{F}}(d) =& \left|\Big\{d'\in\mathcal{D} : \big|\{\gamma\in\Gamma : rank_\gamma(d')\geq rank_\gamma(d)\}\big| \geq |\Gamma|/2\Big\}\right| \\ &\propto P\Big(\big\{d'\in\mathcal{D} : |\{\gamma\in\Gamma : rank_\gamma(d')\geq rank_\gamma(d)\}| \geq |\Gamma|/2\big\}\Big) \\ &\propto -log\left(P\Big(\big\{d'\in\mathcal{D} : |\{\gamma\in\Gamma : rank_\gamma(d')\leq rank_\gamma(d)\}| \geq 1/2\big\}\Big)\right) . \end{aligned}$$

In summary, averaging schemes, the Copeland method, the Borda count algorithm and the UIR are all closely related to different approaches to the Observational Information Quantity that satisfy different desirable properties. In the next section, we will empirically study the suitability of these approaches based on their formal properties and the ranking fusion context (the numbers of rankings and items).

7.5. Empirical Comparison of Ranking Fusion Functions

We present experiments on test collections corresponding to six tasks. All of them are related to some kind of document similarity. The experiments consist of ranking document pairs by means of different similarity prediction systems. Highly similar document pairs should be located at the top. We apply the analysed ranking fusion algorithms to merge all rankings. The selected tasks are described below.

Document Clustering (CL): We use the WePS-1 data set [13], which contains approximately six thousand manually grouped web pages. Here, we consider a set of 167 similarity measures introduced in [12], that employ a wide range of features (from n-grams to different classes of named entities) and provide state-of-the-art results.

Semantic Textual Similarity (STS): We employ the dataset obtained from the pilot task in SemEval-2012 [2], which includes 3050 similarity instances distributed among four sets, and 88 runs (similarity measures). The similarity of pairs of sentences was rated on a 0-5 scale (from low to high similarity) by human judges using Amazon Mechanical Turk.

Textual Entailment (TE): We use the training set provided as part of the RTE-2 evaluation campaign [17], which consists of 800 text-hypothesis pairs. We have developed 102 similarity measures for this scenario (all of them are based on [76]). They all measure word overlap over different text components: levels in the parse tree, PoS tags, lemmas and relations. To preserve the formal properties of these similarity measures, when sentences do not include a text component (e.g. a certain PoS tag), the corresponding similarity is set to 0.5.

Document Retrieval (IR): We use queries 701 to 750 in the GOV-2 collection used in the TREC 2004 Terabyte Track. The document-query similarity measures consist of the outcomes of 60 retrieval systems developed by the participants in the track. We consider the top 100 documents from the output of each search engine.

Machine Translation Evaluation (MT): We use data sets from the Arabic-to-English (AE) and Chinese-to-English (CE) NIST MT Evaluation campaigns in 2004 and 2005[5]. We take the sum of adequacy and fluency, both on a 1-5 scale, as a global manual assessment of quality [71]. These data sets include approximately 8000 similarity instances between MT outputs and human-generated references. As similarity measures, we use 64 automatic evaluation measures provided by the ASIYA Toolkit [47][6]. This set includes measures operating at different linguistic levels (lexical, syntactic, and semantic) and includes all popular measures (BLEU, NIST, GTM, METEOR, ROUGE, etc.)

Automatic Summarization Evaluation (AS): We use the DUC 2005/2006 test collections [7] [29, 30]. At DUC, summaries were evaluated according to several criteria; here, we focus on responsiveness judgements, for which the quality score is an integer between 1 and 5. We employ standard variants of ROUGE [75] as similarity measures.

Using these datasets, we test the (comparative) ability of combined rankings to predict the true similarity of documents. In our experiments, we consider pairs $((d_1, d_2), (d'_1, d'_2))$ of similarity instances. For all of them, there is some difference in similarity according to humans judgement ($sim(d_1, d_2) > sim(d'_1, d'_2)$). We randomly select $10,000$ pairs of similarity instances from each data set.

We test the ability of each method to combine rankings and predict the closest document similarity. The *effectiveness*, Eff($rank_{\mathcal{F}}$), is computed as $P\big(rank_{\mathcal{F}}(d_1, d_2) > rank_{\mathcal{F}}(d'_1, d'_2) \mid sim(d_1, d_2) > sim(d'_1, d'_2)\big)$. When the evaluated method returns the same value for both instances, we estimate effectiveness as 0.5. We normalize the individual rankings to values between 0 and 1 for averaging schemes. We also compare the results with the best and worst rankings from the whole set.

[5] http://www.nist.gov/speech/tests/mt
[6] http://www.lsi.upc.edu/ nlp/Asiya
[7] http://duc.nist.gov/

Table 7.2: Empirical Comparison of Unsupervised Ranking Fusion

	Summarization	**Retrieval**	**Entailment**	**Machine Trans.**	**Semantic Sim.**	**Clustering**
Combining all rankings						
Best measure	0.73	0.69	0.62	0.69	0.74	0.81
Worst measure	0.63	0.50	0.46	0.54	0.52	0.50
UIR	0,70	0,66	0,60	0,65	0,67	0,79
Coppeland	**0,72**	**0,67**	**0,67**	**0,69**	**0,75**	**0,83**
Borda	**0,72**	**0,67**	0,66	0,68	**0,75**	**0,83**
OIQ_{ind}	0,71	**0,67**	0,65	0,68	**0,75**	**0,83**
MinRank	0,67	0,61	0,56	0,64	0,68	0,79
Avg.	0,68	0,66	0,68	0,68	0,71	0,83
Geo.	0,50	0,54	0,51	0,52	0,50	0,50
Harm.	0,68	0,66	0,63	0,62	0,63	0,80
Combining two random rankings						
Best measure	0.706	0.619	0.539	0.654	0.727	0.607
Alternative measure	0.679	0.559	0.505	0.613	0.671	0.529
UIR	**0.703**	**0.613**	**0.539**	**0.654**	**0.723**	**0.617**
Coppeland	**0.703**	**0.613**	**0.539**	**0.654**	**0.723**	**0.617**
Borda	**0.703**	0.612	**0.539**	**0.654**	**0.723**	0.616
OIQ_{ind}	0.700	0.611	0.538	**0.654**	0.720	0.616
MinRank	0.693	0.604	0.532	0.645	0.710	0.615
Avg	0.696	0.606	0.539	0.653	0.723	0.616
Geo	0.693	0.607	0.539	0.647	0.716	0.616
Harm	0.692	0.587	0.539	0.644	0.712	0.616
Max	0.696	0.603	0.513	0.648	0.716	0.612
Min	0.688	0.581	0.537	0.624	0.698	0.554
Combining two random rankings plus five redundant rankings						
Best measure	0.706	0.619	0.539	0.654	0.727	0.607
Alternative measure	0.679	0.559	0.505	0.613	0.671	0.529
UIR	**0.703**	**0.613**	**0.539**	**0.654**	**0.723**	**0.617**
Coppeland	0.681	0.599	0.532	0.631	0.714	0.618
Borda	0.690	0.599	0.536	0.637	0.696	0.615
OIQ_{ind}	0.689	0.598	0.536	0.638	0.698	0.615
MinRank	0.693	0.604	0.532	0.645	0.710	0.615
Avg.	0.687	0.603	0.537	0.638	0.697	0.615
Geo.	0.686	0.603	0.538	0.635	0.695	0.615
Harm.	0.686	0.582	0.538	0.636	0.696	0.615
Max	0.696	0.603	0.513	0.648	0.716	0.612
Min	0.688	0.581	0.537	0.624	0.698	0.554

We consider three experiments. In the first one, all available rankings are merged. In the second, experiment we test the performance of the fusion methods over two rankings only to see how these methods perform with a small number of individual rankings. In the third experiment, we replicate the less predictive ranking 5 times to test the ability of the fusion functions to accommodate redundant rankings without introducing bias.

Table 7.2 shows the results. A salient observation is that in all experiments, fusion methods without any correspondence with the OIQ (MinRank, Max, Min., Harm. and Geo.) achieve lower results than the other methods. In addition, on every dataset, there is at least one method that is able to achieve results similar to the best individual ranking in the combination, corroborating the phenomena observed in the literature.

When combining all rankings, the UIR is less reliable than other methods, although it satisfies *dependence* and *cancellation*. The reason is that the need for sample instances to compute the UIR grows exponentially with the number of rankings when computing unanimous improvements. On the other hand, when only two rankings are combined, the performance of the UIR drastically improves, achieving the best results on two datasets. Moreover, when redundant rankings are added to the set, the UIR is the best fusion method for all datasets because it is not affected by redundancy. Based on these findings, we can conclude that when many rankings are to be combined, monotonicity is an important property to consider.

In the absence of redundant rankings, regardless of whether all or only two rankings are combined, Copeland's method is the best performer across all datasets, indicating that its compromise between dependence and monotonicity is suitable in these situations. In particular, the Copeland results are equivalent to the UIR results when combining two rankings. However, with the addition of redundancy its effectiveness decreases.

Finally, in general, the Borda and OIQ_{ind} methods achieve similar performance, slightly lower than that of the Copeland method and much lower than that of the UIR in the presence of redundancy. This suggests an empirical convergence between the traditional Borda algorithm and the OIQ estimation under the assumption of independence.

7.6. Conclusions: The OIQ as a Formally Grounded Ranking Fusion Algorithm

We have presented an in-depth formal and empirical comparison of unsupervised ranking fusion approaches. Our formal analysis suggests that some conflict exists between capturing ranking monotonicity and dependence. That is, traditional voting approaches such as the Borda, score averaging and Reciprocal Rank Fusion methods are able to capture slight changes in rankings. However, they do not compensate for the effect of redundant rankings. Whereas individual voters have an intrinsic decision weight in the social sciences, in most information access problems, redundant information does not contribute to the task of interest. For instance, in information retrieval, there is no reason to reward documents that are relevant according to multiple identical search engines. On the other hand, approaches such as the UIR, MinRank, minimum and maximum scores satisfy dependence at the cost of monotonicity.

We have seen that, when document ranking scores are considered as features, the OIQ behaves as a fusion ranking method and satisfies (at the theoretical level) the three desirable properties identified in this chapter. However, the OIQ treats an infinite

countable item set as a probabilistic sample space. Thus, we encounter the same conflict (monotonicity vs. dependency) when approaching the OIQ under different statistical assumptions.

On the other hand, we have seen that the OIQ is closely related to the traditional Borda algorithm under the assumption of independence, to the Unanimous Improvement Ratio (UIR) when estimating dependencies, and to the Copeland voting algorithm. As our formal analysis predicts, our empirical study shows that the Copeland approach, which establishes a compromise between dependency and monotonicity achieves competitive results. However, with the addition of redundant rankings, the UIR clearly outperforms the rest of the approaches.

This study offers evidence for the Observational Information Quantity as a theoretical foundation of ranking fusion algorithms. This formal perspective allows us to identify the bottleneck in ranking fusion for future works, i.e., the estimation of the feature distributions.

8

Conclusions

Document representation is especially challenging in unsupervised scenarios. In the absence of reference outputs, it is not easy to determine the feature weights, dependencies and scaling. To overcome these challenges, we have presented a formal framework for document representation.

We have identified three main properties that a representation framework should possess: (i) *specificity*, which establishes that the less frequently a feature appears, the more relevant it is, since it more effectively distinguishes an object from others, (ii) *dependence*, which establishes that redundant features do not provide information; and (iii) *quantitativity*, the need to capture both discrete and quantitative feature values.

Based on an analysis of the literature (Chapter 2), one of the first conclusions drawn in this thesis is that Shannon's information content ($IC(x) = -log(Px)$) is an underlying notion in many representation frameworks that captures specificity and dependence. It is related to tf·idf in the BoW approach, perplexity in language models and information-theory-based variants of feature set representations. Another conclusion is that a conflict exists between feature value quantitativity and specificity (and with feature dependence). Most existing representation frameworks capture specificity and, in some cases dependence, but not quantitativity. To capture quantitativity, documents have been represented as vectors and fuzzy sets. However, in that case, dependence and specificity are ignored. The underlying reason for this conflict is that Shannon's information content is defined for discrete events. Note that Shannon based his theory on discrete notions such as yes/no questions and bits. In representation, this translates into discrete text features such as words or n-grams and their occurrence. The problem is that it is not possible to estimate the probability of continuous feature values. Therefore, the notion of information content cannot be applied.

The Observational Representation Framework (ORF) presented in this thesis (see Chapter 3) attempts to overcome this challenge under the following assumptions:

- While the probability sample spaces in previous representation frameworks have consisted of features (words, n-grams, etc.), in the ORF the sample space is made up of the infinite and countable universe of documents. In other words, each

document is a probabilistic event in our probabilistic framework. Note that, with an infinite document space, this probability tends towards zero.

- To prevent the management of zero probability events, our framework distinguishes between a document itself and the corresponding *observation outcome*. An observation outcome is an instantiation of feature values, for instance, the occurrence of the words "Once", "upon", "a", and "time". In other words, this observation is subsumed by every story containing these words. The likelihood of an observation outcome is given by the mass probability of the documents subsuming it, which we call the Observational Information Quantity (OIQ).
- Feature quantitativity is captured modelling observation outcomes as fuzzy sets instead of crisp sets. Then, the subsumption of observation outcomes in documents enables the derivation of cumulative probabilities of feature values. That is, the probability of a feature value, e.g. $f_i(d) = 0.3$, is the cumulative probability of values above 0.3 in the document sample set.

We can highlight several interesting implications of the ORF and OIQ. The ORF not only possesses the three properties highlighted in the previous chapter (specificity, dependence and quantitativity), but also exhibits other interesting properties, such as monotonicity with respect to feature values, feature sets and the union of observation outcomes. In addition, the OIQ generalizes Shannon's notions regarding an information quantity as well as the most popular representation methods, such as *tf*, *tf·idf* and perplexity in language models, by assuming a convenient hypothesis. To the best of our knowledge, the OIQ is the first representation model that captures the informativeness (specificity) of quantitative features.

In this thesis, we have explored the strengths of the proposed representation framework in various ways. More specifically, we have studied the ability of the ORF and OIQ to:

- Define a general axiomatic framework for similarity functions based on metric spaces, feature set operators, information theory and probabilistic events.
- Integrate, in the same representation, intrinsic features such as words, with extrinsic features such as clustering outputs or class memberships.
- Behave as an unsupervised document ranking fusion algorithm.

8.1. Modelling Similarity

Our analysis suggests that the properties of similarity functions are closely related to the corresponding representation frameworks and their properties (see Chapter 4), such as the properties of metric spaces (maximality, triangular inequality and symmetry) or Tversky's Feature Contrast Model (matching, monotonicity and independence). Our counterexamples and previous studies in cognitive science indicate that existing geometric and set-based axiomatic do not capture every aspects of similarity in information access scenarios.

On the other hand, the generality of the ORF opens the door for us to define a general framework for comparing diverse similarity functions. This framework consists of a set of formal constraints that capture aspects of similarity functions based on metric spaces, feature sets, information theory, and probabilistic events. These constraints are as follows: (i) *identity*, which states that adding or removing characteristics to or from a pair of instances decreases their similarity; (ii) *identity specificity*, which states that more specific pieces of information are more informative; (iii) *unexpectedness*, which establishes that less common characteristics should have greater importance; and (iv) *dependency*, which establishes the existence of relationships between characteristics. It is also possible to consider one more constraint: (v) *asymmetricity*, which establishes that the order in which two documents are compared affects the human notion of similarity.

From an analysis of the literature on similarity functions, we can highlight the following two conclusions (Chapter 5). First, none of the reviewed similarity functions satisfies every constraint simultaneously, but in general, their weaknesses are mitigated at the document representation level; for example, specificity is addressed by means of *tf*, *tf·idf*, unexpectedness is addressed through smoothing in language models, and the dependence constraint is addressed via dimensionality reduction methods. The second conclusion is that the pointwise mutual information (PMI) and conditional probability can be understood as the two basic dimensions of similarity. Together, they satisfy every constraint and a single property called *similarity information monotonocity* (SIM) can be identified that subsumes the four first constraints and is defined in terms of the PMI and conditional probability.

Based on the ORF and SIM, we have defined a similarity function called the *Information Contrast Model* (ICM), as introduced in Chapter 5. The ICM is a parametrized generalization of the PMI. To our knowledge, it is the only similarity function that satisfies every formal constraint, whenever the parameters are fixed in a certain range. The ICM converges to the PMI and conditional probability for extreme parameter values. In addition, the ICM generalizes the set-based similarity functions. A small empirical case study over image descriptors shows that the ICM overcomes the shortcomings observed in similarity counterexamples used in the analysis of previous axiomatic.

8.2. Aggregating Heterogeneous Features

The properties of the ORF allow documents to be characterized by both their intrinsic features (words, n-grams) and extrinsic features such as outputs from unsupervised algorithms (e.g., clustering assignments) and supervised algorithms (e.g., class membership). In addition, the OIQ allows all of these features to be combined in the form of an information quantity. We have studied the practical consequences in the context of reputation-monitoring (Chapter 6). In our experiments, the goal was to predict whether two tweets referred to the same subtopic when analysing the reputation of companies on Twitter. For this purpose, we represented tweets by means of word features together with the output of a Bayesian classifier (supervised signal) acting on reputation categories and known subtopics as well as the output of tweet clustering (unsupervised

signal). These experiments demonstrated that progressively adding features increases the performance of the tweet similarity computation. This result was verified under various similarity functions (the pointwise mutual information, the Jaccard and Lin distances and the Information Contrast Model).

8.3. The ORF and OIQ in Ranking Fusion

The third application of the ORF studied in this thesis is a demonstration that when document rank scores are considered as features, the OIQ behaves as an unsupervised ranking fusion algorithm (Chapter 7). At the theoretical level, it exhibits three properties that are not possessed by traditional approaches, such as Reciprocal Rank Fusion, score averaging schemes or voting algorithms. These three properties are (i) *monotonicity*, which states that items with high scores in all rankings should have a high score in the combined ranking; (ii) *dependence*, which states that the combined results show better performance if redundancies are not considered; and (iii) *cancellation*, which indicates the way in which two exactly opposite rankings should be combined.

However, in practice, the limitations of probability estimation prevent the OIQ from satisfying, all these properties simultaneously due to the size of the document collection, computational cost considerations and the combinatorial explosion of the number of pairs of ranking positions. Depending on the assumptions adopted for estimation purposes, some properties are sacrificed. Our formal analysis shows that depending on the approach taken, the OIQ is closely related to different ranking fusion algorithms. Our empirical results confirm that different properties are more determinant of the effectiveness of ranking fusion than others in different situations.

8.4. Limitations and Future Work

The main limitations of the proposed representation framework originate from the assumptions adopted at the empirical level, which are basically the assumptions of *dependency* and *additivity*. Alternative representation models have relied on various approaches for addressing *dependency* (reduction of dimensionality, dependency modelling, direct probabilistic calculation, etc.), but all of them are approximate solutions.

We have left several issues to be addressed in future work; among them, we highlight the following:

- In Section 5.2, we analyse similarity as a distance in a metric space; in particular, the cosine similarity distance is one of the most representative measures of this kind, and it would be interesting to analytically verify whether the cosine similarity possesses the DEPENDENCY property.
- In Chapter 5, a new similarity function (ICM) is proposed, which assumes that the statistical dependencies of pieces of information may be determined by the

user context. It is left to future studies to empirically verify how the ICM can be combined with appropriate representation spaces to yield operational similarity models.

- In Section 6.7, the proposed model (OIQ) is parametrized to improve the performance of similarity estimation. The parametrized model studied here is linear in its components, and the optimization of verify the proposed model with non-linear kernels remains to be investigated.

- As noted in Chapter 3, one of the main limitations of the proposed model is the dependency assumption at the empirical level. We believe that it could be possible to model the dependence directly with features, for instance, by positing a formal model in which the dependencies are drawn from the syntactical structure of the sentences.

List of Publications

- Damiano Spina, Jorge Carrillo-de-Albornoz, Tamara Martín, Enrique Amigó, Julio Gonzalo and **Fernando Giner**. *UNED Online Reputation Monitoring Team at RebLab 2013*. In CLEF 2013 Labs and Workshops Notebook Papers (2013). GII-GRIN-SCIE Conference Rating (GGS) Class 3.
- **Fernando Giner** and Enrique Amigó. *General Representation Model for Text Similarity*. In International Workshop on Future and Emerging Trends in Language Technology. Pages 158–169, 2016, Springer.
- Enrique Amigó, **Fernando Giner**, Julio Gonzalo and Felisa Verdejo. *A Formal and Empirical Study of Unsupervised Signal Combination For Textual Similarity Tasks*. In European Conference on Information Retrieval (ECIR'2017). Pages 369–382, 2017, Springer. GII-GRIN-SCIE Conference Rating (GGS) Class 2.
- Enrique Amigó, **Fernando Giner**, Julio Gonzalo and Felisa Verdejo. *An Axiomatic Account of Similarity*. In Proceedings of the SIGIR'17 Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks (ATIR), 2017. GII-GRIN-SCIE Conference Rating (GGS) Class 1.
- Enrique Amigó, **Fernando Giner**, Stefano Mizzaro and Damiano Spina. *A Formal Account of Effectiveness Evaluation and Ranking Fusion*. In Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval. Pages 123–130, 2018, ACM. GII-GRIN-SCIE Conference Rating (GGS) Class 1.
- **Fernando Giner**, Enrique Amigó and Felisa Verdejo. *Integrating Learned and Explicit Document Features for Reputation Monitoring in Social Media*. Journal: Knowledge and Information Systems. Volume 62, n. 3, pages 951–985, 2020, Springer. Journal Impact Factor (JCR): 2936.
- Enrique Amigó, **Fernando Giner**, Julio Gonzalo and Felisa Verdejo. *On the Foundations of Similarity in Information Access*. Journal: Information Retrieval Journal, Volume 23, n. 3, pages 216–254, 2020. Journal Impact Factor (JCR): 2209.

A

Formal Proofs

A.1. Formal Proofs for Chapter 3

Proposition 3.1. *The OIQ of a union of observation outcomes is equivalent to the OIQ of the maximum feature values. Formally, given two observation outcomes X and Y:*

$$\mathcal{I}\big(X \cup Y\big) = -log\left(P\Big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq \ \max(x_i, y_i),\ i = 1, \ldots, |\Gamma|\}\Big)\right) \ .$$

Proof A.1.1. The proof is straightfordward. According to the fuzzy set operators:

$$X \cup Y = (\Gamma, f) \ ,$$

where

$$f(\gamma_i) = \max(x_i, y_i),\ i = 1, \ldots, |\Gamma| \ .$$

□

Property 3.1. *Increasing the feature instantiation values increases the OIQ. Given two observation outcomes (fuzzy feature sets) X and Y the following is verified:*

$$x_i \geq y_i, \quad \forall\ \gamma_i \in \Gamma \Rightarrow \mathcal{I}\big(X\big) \geq \mathcal{I}\big(Y\big) \ .$$

Proof A.1.2. From $x_i \geq y_i, \quad \forall\ \gamma_i \in \Gamma$, it follows,

$$\Big\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq x_i,\ \forall\ \gamma_i \in \Gamma\Big\} \subseteq \Big\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq y_i,\ \forall\ \gamma_i \in \Gamma\Big\} \ .$$

Then,

$$P\Big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq x_i,\ \forall\ \gamma_i \in \Gamma\}\Big) \leq P\Big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq y_i,\ \forall\ \gamma_i \in \Gamma\}\Big) \ .$$

This implies that:

$$P\Big(\{d \in \mathcal{D} : \mathcal{O}_\Gamma(d) \supseteq X\}\Big) \leq P\Big(\{d \in \mathcal{D} : \mathcal{O}_\Gamma(d) \supseteq Y\}\Big) \ .$$

And therefore, according to definition 3.4:

$$\mathcal{I}(X) \geq \mathcal{I}(Y) \ .$$

□

Property 3.2. *Adding features to the set* Γ *increases the OIQ values of document observation outcomes. Let* X *and* X_{sub} *be two observation outcomes such that* $X = (\Gamma, f)$ *and* $X_{sub} = (\Gamma - \{\gamma\}, f)$*:*

$$\mathcal{I}(X) \geq \mathcal{I}(X_{sub}) \ .$$

Proof A.1.3. Notice that if we add a feature, the new document observation outcome is more restrictive than the initial observation, and thus, the set of documents which verify the new observation is contained in the set of documents which verify the initial observation, $\mathcal{O}_{\Gamma\cup\{\gamma'\}}(d) \subseteq \mathcal{O}_{\Gamma}(d)$. Then,

$$P\Big(\{d \in \mathcal{D} : \mathcal{O}_{\Gamma}(d) \supseteq X\}\Big) \leq P\Big(\{d \in \mathcal{D} : \mathcal{O}_{\Gamma-\{\gamma\}}(d) \supseteq X_{sub}\Big)$$

And therefore, according to definition 3.4:

$$\mathcal{I}(X) \geq \mathcal{I}(X_{sub}) \ .$$

□

Property 3.3. *Given two observation outcomes* X *and* Y*, the OIQ of their union is larger than those of the individual outcomes:*

$$\mathcal{I}(X \cup Y) \geq \mathcal{I}(X) \ .$$

Proof A.1.4. Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ be, the corresponding representations of two observation outcomes, by Proposition 3.1:

$$\mathcal{I}(X \cup Y) = -log\left(P\Big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq \ \max(x_i, y_i),\ i = 1, \ldots, |\Gamma|\}\Big)\right) \ .$$

Given that,

$$P\Big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq \ \max(x_i, y_i),\ \forall\ \gamma_i \in \Gamma\}\Big) \leq P\Big(\{d \in \mathcal{D} : \pi_d(\gamma_i) \geq \ x_i,\ \forall\ \gamma_i \in \Gamma\}\Big) \ .$$

We finally get, $\mathcal{I}(X \cup Y) \geq \mathcal{I}(X)$. Similarly, we can get the same result for Y. □

Property 3.4. *The more unexpected a feature instantiation is, the more informative it is. Given two single feature observation outcomes* $X = (\{\gamma\}, f)$ *and* $X' = (\{\gamma'\}, f')$ *where* $f(\gamma) = f'(\gamma') = v$*, we have the following implication:*

$$P(\{d \in \mathcal{D} : \pi_d(\gamma) \geq v\}) < P(\{d \in \mathcal{D} : \pi_d(\gamma') \geq v\}) \Longrightarrow \mathcal{I}(X) \geq \mathcal{I}(X') \ .$$

Proof A.1.5. By hypothesis,

$$P\big(\{d \in \mathcal{D} : \pi_d(\gamma) \geq f(\gamma)\}\big) \leq P\big(\{d \in \mathcal{D} : \pi_d(\gamma') \geq f'(\gamma')\big) \ .$$

Which is equivalent to,

$$\frac{1}{P\big(\{d \in \mathcal{D} : \pi_d(\gamma) \geq f(\gamma)\}\big)} \geq \frac{1}{P\big(\{d \in \mathcal{D} : \pi_d(\gamma') \geq f'(\gamma')\big)} \Rightarrow$$

$$\log\left(\frac{1}{P\big(\{d \in \mathcal{D} : \pi_d(\gamma) \geq f(\gamma)\}\big)}\right) \geq \log\left(\frac{1}{P\big(\{d \in \mathcal{D} : \pi_d(\gamma') \geq f'(\gamma')\big)}\right) \Rightarrow$$

$$\Rightarrow \mathcal{I}(X) \geq \mathcal{I}(X') \ .$$

□

Property 3.5. *Redundant features do not affect the OIQ values of document observation outcomes. Given two features γ_1, $\gamma_2 \in \Gamma$, if there exists a real strict monotonic function g, that satisfies: $\pi_d(\gamma_1) = g\big(\pi_d(\gamma_2)\big)$, $\forall\ d \in \mathcal{D}$, then, $\mathcal{I}_{\{\gamma_1\}}(d) = \mathcal{I}_{\{\gamma_1,\gamma_2\}}(d)$, $\forall\ d \in \mathcal{D}$.*

Proof A.1.6. Consider two features, γ_1, $\gamma_2 \in \Gamma$, given a document, $d \in \mathcal{D}$, it produces an observation outcome under γ_1, $\mathcal{O}_{\{\gamma_1\}}(d)$, whose OIQ is:

$$\mathcal{I}_{\{\gamma_1\}}(d) = -log\Big(P\big(\{d' \in \mathcal{D} : \pi_{d'}(\gamma_1) \geq \pi_d(\gamma_1)\}\big)\Big) =$$

$$= -log\Big(P\Big(\{d' \in \mathcal{D} : g\big(\pi_{d'}(\gamma_2)\big) \geq g\big(\pi_d(\gamma_2)\}\big)\Big)\Big) \ .$$

Given that g is a strict monotonic function,

$$-log\Big(P\Big(\big\{d' \in \mathcal{D} : g(\pi_{d'}(\gamma_2)) \geq g(\pi_d(\gamma_2))\big\}\Big)\Big) =$$

$$= -log\Big(P\Big(\big\{d' \in \mathcal{D} : g(\pi_{d'}(\gamma_2)) \geq g(\pi_d(\gamma_2)), \pi_{d'}(\gamma_2) \geq \pi_d(\gamma_2)\big\}\Big)\Big) = \mathcal{I}_{\{\gamma_1,\gamma_2\}}(d) \ .$$

□

Property 3.6. *The OIQ of a document observation outcome under an infinite number of heterogeneous features corresponds to the likelihood of the document itself.*

$$\lim_{|\Gamma|\to\infty} \mathcal{I}_\Gamma(d) = -\log\big(P(d)\big) \ .$$

Proof A.1.7. Assume that we have a finite set of documents, the proof of this proposition is a direct consequence of the representativity of the documents by the features. If we have an infinite set of features, then they will describe every document, and documents will be unequivocally determined by the values of a set of features. □

Property 3.7. *The OIQ for two inverse and continuous-valued features corresponds to the probability of equality in this feature. Given a document $d \in \mathcal{D}$ and a feature $\gamma \in \Gamma$, consider the set of features $\{\gamma, \gamma^{-1}\}$, where γ^{-1} is defined by $\pi_d(\gamma^{-1}) = \pi_d(\gamma)^{-1}$, $\forall d \in \mathcal{D}$; then:*

$$\mathcal{I}_{\{\gamma,\gamma^{-1}\}}(d) = -\log\Big(P\big(\{d' : \pi_{d'}(\gamma) = \pi_d(\gamma)\}\big)\Big) \ .$$

Proof A.1.8. Given a fixed document, $d \in \mathcal{D}$, consider all the documents, $d' \in \mathcal{D}$, which verify the inequalities:

$$\pi_{d'}(\gamma) \leq \pi_d(\gamma) \ \wedge \ \pi_{d'}(\gamma^{-1}) \leq \pi_d(\gamma^{-1}) \ .$$

These inequalities are equivalent to (by definition of γ^{-1}):

$$\pi_{d'}(\gamma) \leq \pi_d(\gamma) \ \wedge \ \frac{1}{\pi_{d'}(\gamma)} \leq \frac{1}{\pi_d(\gamma)} \ .$$

Notice that $\pi_d(\gamma)$ and $\pi_{d'}(\gamma)$ are non-negative numbers, therefore, these inequalities imply that: $\pi_d(\gamma) = \pi_{d'}(\gamma)$. Then, the Observational Information Quantity is:

$$\mathcal{I}_{\{\gamma,\gamma^{-1}\}}(d) = -\log\left(P\Big(\big\{d' \in \mathcal{D} : \pi_{d'}(\gamma) \leq \pi_d(\gamma) \wedge \pi_{d'}(\gamma^{-1}) \leq \pi_d(\gamma^{-1})\big\}\Big)\right) .$$

Which is equivalent to:

$$\mathcal{I}_{\{\gamma,\gamma^{-1}\}}(d) = -\log\Big(P\big(\{d' \in \mathcal{D} : \pi_{d'}(\gamma) = \pi_d(\gamma)\}\big)\Big) .$$

□

Proposition 3.2. *Under the assumptions of word information additivity and equiprobability, the OIQ is equivalent to the tf representation. Let $d = (x_1, \ldots, x_n)$ be the tf representation of a document d with respect to the vocabulary $\Gamma = \{\chi_{w_1}, \ldots, \chi_{w_n}\}$:*

$$\mathcal{I}_{\{\chi_{w_i}\}}(d) = tf(w_i, d) = x_i \ .$$

Proof A.1.9. Given the vocabulary, $\mathcal{V} = \{w_1, \ldots, w_n\}$, consider the set of features as $\Gamma = \{\chi_{w_1}, \ldots, \chi_{w_n}\}$, and given a document from the collection, $d \in \mathcal{D}$, we are interested in computing the OIQ: $\mathcal{I}_{\{\chi_{w_i}\}}(d)$.

Assuming information additivity and considering text words as basic linguistic units, we have,

$$\mathcal{I}_{\{\chi_{w_i}\}}(d) = \sum_{w_j \in d} \mathcal{I}_{\{\chi_{w_i}\}}(w_j) = \sum_{w_j \in d} -\log\Big(P\big(\{w' \in \mathcal{V} : \chi_{w_i}(w') \geq \chi_{w_i}(w_j)\}\big)\Big) .$$

Notice that, if $w_j \neq w_i$, then $\chi_{w_i}(w_j) = 0$. Thus, $P\big(\chi_{w_i}(w') \geq 0\big) = 1$, since by definition $\chi_{w_i}(d) \geq 0$, $\forall d \in \mathcal{D}$. Therefore, in the last summation all the terms are null, except for $w_j = w_i$. In this case, we have that $\chi_{w_i}(w_i) = 1$, and given that by definition of the function $\chi_{w_i}(.)$, its maximum value is 1, we can say that $\chi_{w_i}(w') \geq 1$ is equivalent to $\chi_{w_i}(w') = 1$. Therefore, the probability $P\big(\chi_{w_i}(w') = 1\big)$ is exactly $P(w' = w_i) = P(w_i)$. And, $\mathcal{I}_{\{\chi_{w_i}\}}(d) \propto -\log\big(P(w_i)\big)$.

One of the assumptions is that every word is equiprobable, i.e., $P(w_i) = k$, $1 \leq i \leq n$, for an arbitrary k. In order to achieve the result, we can choose k in such a way that $-\log(k) = 1$. And finally, the summation give us the $tf(w_i, d)$. □

Proposition 3.3. *When word occurrences are taken as features, the observational in-*

formation quantity of a word is equivalent to its idf:

$$\mathcal{I}_{\{\chi_w\}}\big(w\big) = idf(w) \ .$$

Proof A.1.10. Consider the set of features as $\Gamma = \{\chi_w\}$, given the document $d = \{w\}$ from the collection, we are interested in computing the OIQ: $\mathcal{I}_{\{\chi_w\}}\big(w\big)$.

By Definition 3.4, we have,

$$\mathcal{I}_{\{\chi_w\}}\big(w\big) = -\log\Big(P\big(\{d' \in \mathcal{D} : \chi_w(d') \geq \chi_w(w)\}\big)\Big) \ .$$

Notice that, $\chi_w(w) = 1$, and given that by definition of the function $\chi_w(.)$, its maximum value is 1, we can say that $\chi_w(d') \geq 1$ is equivalent to $\chi_w(d') = 1$. Therefore, the expression $-\log\Big(P\big(\{d' \in \mathcal{D} : \chi_w(d') = 1\}\big)\Big)$ is exactly:

$$-\log\Big(P\big(\{d' \in \mathcal{D} : w \in d'\}\big)\Big) = idf(w) \ .$$

And thus, $\mathcal{I}_{\{\chi_w\}}\big(w\big) = idf(w)$. □

Proposition 3.4. *When word occurrences are taken as features and under the assumption of information additivity, the OIQs of single features are equivalent to the* tf·idf *representation.*

$$\mathcal{I}_{\{\chi_{w_i}\}}\big(d\big) = tf(w_i, d) \cdot idf(w_i) \ .$$

Proof A.1.11. Given the vocabulary, $\mathcal{V} = \{w_1, \ldots, w_n\}$, considering the set of features as, $\Gamma = \{\chi_{w_1}, \ldots, \chi_{w_n}\}$, and given a document from the collection, $d \in \mathcal{D}$, we are interested in computing the OIQ: $\mathcal{I}_{\{\chi_{w_i}\}}\big(d\big)$.

Assuming information additivity, and considering documents as basic linguistic units, we have,

$$\mathcal{I}_{\{\chi_{w_i}\}}\big(d\big) = \sum_{w_j \in d} \mathcal{I}_{\{\chi_{w_i}\}}\big(w_j\big) = \sum_{w_j \in d} -\log\Big(P\big(\{d' \in \mathcal{D} : \chi_{w_i}(d') \geq \chi_{w_i}(w_j)\}\big)\Big) \ .$$

Notice that, if $w_j \neq w_i$, then $\chi_{w_i}(w_j) = 0$. Thus, $P\big(\{d' \in \mathcal{D} : \chi_{w_i}(d') \geq 0\}\big) = 1$, since by definition $\chi_{w_i}(d') \geq 0$, $\forall d' \in \mathcal{D}$. Therefore, in the last summation all the terms are null, except for $w_j = w_i$. In this case, we have as many terms as the number of times that the word w_i appears in the document d, i.e., $tf(w_i, d)$. Moreover, we have that $\chi_{w_i}(w_i) = 1$, and given that by definition of $\chi_{w_i}(.)$, its maximum value is 1, we can say that $\chi_{w_i}(d') \geq 1$ is equivalent to $\chi_{w_i}(d') = 1$. Therefore, the expression $-\log\Big(P\big(\{d' \in \mathcal{D} : \chi_{w_i}(d') = 1\}\big)\Big)$ is exactly $-\log\Big(P\big(\{d' \in \mathcal{D} : w_i \in d'\}\big)\Big) = idf(w_i)$. And thus, $\mathcal{I}_{\{\chi_{w_i}\}}\big(d\big) = tf(w_i, d) \cdot idf(w_i)$. □

Proposition 3.5. *When the occurences of word-position pairs are taken as features and under the assumption that the documents in the collection are generated from a probability distribution Θ, the perplexity over the language model defined by θ of a word*

sequence $d = (w_1, \ldots, w_m)$*, is an exponential function of the OIQ:*

$$Perplexity(d) = 2^{\frac{1}{m}\mathcal{I}_\Gamma\left(d\right)} \ .$$

Proof A.1.12. Given the vocabulary, $\mathcal{V} = \{w_1, \ldots, w_m\}$, a language model, θ, is defined as a probability distribution:

$$\Theta = \{w_i, P_{lm}(w_i \mid \theta) : w_i \in \mathcal{V}\} \ .$$

With $\sum_{i=1}^{n} P_{lm}(w_i \mid \theta) = 1$.

Given a document, $d = \{w_1, \ldots, w_n\} \in \mathcal{D}$, we can consider as fuzzy feature set, X, the corresponding pairs word-position, $\{(w_1, 1), \ldots, (w_n, n)\}$; where the membership function is binary (the occurence of the feature). Then, by definition of OIQ:

$$\mathcal{I}\big(X\big) = -log\Big(P\big(\{d \in \mathcal{D} : d \ contains \ \{(w_1, 1), \ldots, (w_n, n)\}\}\big)\Big) =$$

$$= -log\Big(P\big(\{d \in \mathcal{D} : d = (w_1, \ldots, w_n)\}\big)\Big) \ .$$

Given that documents are generated from a language model described by the probability distribution P_{lm}, then, the probability that a document is a sequence of words corresponds with the probability of a sequence according to the probability distribution P_{lm}. That is:

$$P\Big(\{d \in \mathcal{D} : d = (w_1, \ldots, w_n)\}\Big) = P_{lm}\Big(d = (w_1, \ldots, w_n)\Big) \ .$$

Then, $\mathcal{I}\big(X\big) = -log\Big(P_{lm}\big(d = (w_1, \ldots, w_n)\big)\Big)$, And finally, with trivial algebraic operations, we have,

$$Perplexity(d) = 2^{\frac{1}{m}\mathcal{I}_\Gamma\left(d\right)} \ .$$

□

Proposition 3.6. *Given two documents* d_1 *and* d_2*, when word occurrences are taken as features and under the assumptions of feature independence and information additivity, Lin's distance can be expressed as:*

$$Lin\big(d_1, d_2\big) = \frac{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1) \cap \mathcal{O}_\Gamma(d_2)\big)}{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1)\big) + \mathcal{I}\big(\mathcal{O}_\Gamma(d_2)\big)} \ .$$

Proof A.1.13. Considering the definition of Lin's distance and assuming information additivity,

$$Lin(d_1, d_2) = \frac{\sum\limits_{w \in d_1 \cap d_2} \mathcal{I}_\Gamma\big(w\big)}{\sum\limits_{w \in d_1} \mathcal{I}_\Gamma\big(w\big) + \sum\limits_{w \in d_2} \mathcal{I}_\Gamma\big(w\big)} \ .$$

Assuming feature independence, it is equivalent to:

$$\frac{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1) \cap \mathcal{O}_\Gamma(d_2)\big)}{\mathcal{I}\big(\mathcal{O}_\Gamma(d_1)\big) + \mathcal{I}\big(\mathcal{O}_\Gamma(d_2)\big)} \ .$$

□

A.2. Formal Proofs for Chapter 4

Proposition 4.1. *The Feature Contrast Model satisfies the metric space Axioms if* $\alpha = 0$ *and* $\beta_1 = \beta_2$*, i.e.,* $Sim(X,Y) = -\beta \cdot f(X \setminus Y) - \beta \cdot f(Y \setminus X)$*, and the salience function is additive for disjoint documents:*

$$X \cap Y = \emptyset \implies f(X \cup Y) = f(X) + f(Y) \ . \tag{A.1}$$

Proof A.2.1. Maximality is satisfied given that:

$$Sim(X,X) = -\beta_1 \cdot f(X \setminus X) - \beta_2 \cdot f(X \setminus X) = 0 - 0 = 0 \ .$$

In addition the salience function f is always positive. Therefore:

$$Sim(X,Y) = -\beta_1 \cdot f(X \setminus Y) - \beta_2 \cdot f(Y \setminus X) \leq 0 \ .$$

The proof for symmetricity is straightfordward, given that $\beta_1 = \beta_2$. Let us prove that it satisfies Triangular Inequality. According to the properties of set operators, we can state that $f(X) = f((X \setminus Y) \cup (X \cap Y))$. And according to Equation A.1:

$$f((X \setminus Y) \cup (X \cap Y)) = f(X \setminus Y) + f(X \cap Y) \implies f(X \setminus Y) = f(X) - f(X \cap Y) \ . \tag{A.2}$$

Then, we need to prove that:

$$sim(X,Y) = -\beta_1 \cdot f(X \setminus Y) - \beta_2 \cdot f(Y \setminus X) \geq$$
$$sim(X,Z) + sim(Z,Y) = -\beta_1 \cdot f(X \setminus Z) - \beta_2 \cdot f(Z \setminus X) - \beta_1 \cdot f(Z \setminus Y) - \beta_2 \cdot f(Y \setminus Z) \ .$$

Given that $\beta_1 = \beta_2$, what we need to prove is:

$$-f(X \setminus Y) - f(Y \setminus X) \geq -f(X \setminus Z) - f(Z \setminus X) - f(Z \setminus Y) - f(Y \setminus Z) \iff$$
$$f(X \setminus Y) + f(Y \setminus X) \leq f(X \setminus Z) + f(Z \setminus X) + f(Z \setminus Y) + f(Y \setminus Z) \ .$$

According to Equation A.2:

$$\begin{aligned} f(X) - f(X \cap Y) + f(Y) - f(Y \cap X) \leq f(X) - f(X \cap Z) + f(Z) - f(Z \cap X) \\ + f(Z) - f(Z \cap Y) + f(Y) - f(Y \cap Z) \iff \\ -2 \cdot f(X \cap Y) \leq 2 \cdot f(Z) - 2 \cdot f(X \cap Z) - 2 \cdot f(Z \cap Y) \iff \\ -f(X \cap Y) \leq f(Z) - f(X \cap Z) - f(Z \cap Y) \iff \\ f(X \cap Y) \geq f(X \cap Z) + f(Z \cap Y) - f(Z) \ . \end{aligned}$$

Now, let as consider the following disjoint subsets:

$$A = X \cap Y \cap Z, \quad B = (X \cap Y) \setminus A, \quad C = (X \cap Z) \setminus A,$$

$$D = (Y \cap Z) \setminus A, \quad E = Z \setminus (X \cup Y) \ .$$

We can state that, $X \cap Y = A \cup B$, $X \cap Z = A \cup C$, $Y \cap Z = A \cup D$ and $Z = A \cup C \cup D \cup E$. Now, according to Equation A.1:

$$f(X \cap Y) \geq f(X \cap Z) + f(Z \cap Y) - f(Z) \iff$$

$$f(A)+f(B) \geq f(C)+f(A)+f(A)+f(D)-f(A)-f(D)-f(C)-f(E) \iff f(B) \geq -f(E) \ .$$

which is true, given that the salience function are necessarily positive. □

Proposition 4.2. *The Tversky's monotonicity axiom is not compatible with the* DEPENDENCY *constraint.*

Proof A.2.2. Let us consider the situation in which X, Y, Z and Z' are four non empty disjoint sets of features, and in addition, the DEPENDENCY conditions hold: $P(XZ \mid YZ') < P(X \mid Y)$ and $P(YZ \mid XZ) < P(Y \mid X)$. Then, according to the DEPENDENCY constraint:

$$Sim(X, Y) < Sim(XZ, YZ') \ .$$

On the other hand, at least when the sets are not overlapped, we can also assert that:

$$\begin{aligned} X \cap Y = XZ \cap YZ' \\ XZ \setminus YZ' = XZ \setminus Y \supseteq X \setminus Y \\ YZ' \setminus XZ = YZ \setminus X \supseteq X \setminus Y \ . \end{aligned}$$

Therefore, accoring to Tverski's monotonocity axiom:

$$Sim(X, Y) \geq Sim(XZ, YZ') \ .$$

which contradicts the DEPENDENCY constraint. □

Lemma 4.1. *SIM is equivalent to stating that a positive similarity increase occurs when both the information quantities of the compared fuzzy feature sets and their sum increase to a greater extent than the information quantity of the union of the fuzzy*

feature sets:

$$\Delta\mathcal{I}(X) + \Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \iff \Delta PMI(X,Y) \geq 0$$

$$\text{and } \Delta\mathcal{I}(X) \geq \Delta\mathcal{I}(X \cup Y) \iff \Delta P(X \mid Y) \geq 0$$

$$\text{and } \Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \iff \Delta P(Y \mid X) \geq 0 \ .$$

Proof A.2.3. First, notice that according to our probabilistic framework, the occurrence of a union of two fuzzy feature sets, X and Y, considers all the documents which have a less projection of feature values than the union, $X \cup Y$, i.e., it considers all the documents which have a less projection of feature values than the occurrence of X, and those which have a less projection of feature values than the occurrence of Y.

Therefore:

$$\mathcal{I}(X \cup Y) = -log\Big(P\big(Occ(X \cup Y)\big)\Big) = -log\Big(P\big(Occ(X), Occ(Y)\big)\Big) \ ,$$

which is denoted in this thesis as $-log(P(X,Y))$. Then:

$$\begin{aligned}
&\Delta\mathcal{I}(X) + \Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \iff \\
&log\left(\frac{1}{P(X')}\right) + log\left(\frac{1}{P(Y')}\right) - \left(log\left(\frac{1}{P(X)}\right) + log\left(\frac{1}{P(Y)}\right)\right) \geq \\
&log\left(\frac{1}{P(X',Y')}\right) - log\left(\frac{1}{P(X,Y)}\right) \iff \\
&log\left(\frac{P(X)\cdot P(Y)}{P(X')\cdot P(Y')}\right) \geq log\left(\frac{P(X,Y)}{P(X',Y')}\right) \iff \\
&\frac{P(X)\cdot P(Y)}{P(X')\cdot P(Y')} \geq \frac{P(X,Y)}{P(X',Y')} \iff \frac{P(X',Y')}{P(X')\cdot P(Y')} \geq \frac{P(X,Y)}{P(X)\cdot P(Y)} \iff \\
&PMI(X',Y') \geq PMI(X,Y) \iff \Delta PMI(X,Y) \geq 0 \ .
\end{aligned}$$

The other two conditions are also equivalent; $\Delta\mathcal{I}(X) \geq \Delta\mathcal{I}(X \cup Y)$ is equivalent to:

$$\begin{aligned}
&log\left(\frac{1}{P(X')}\right) - log\left(\frac{1}{P(X)}\right) \geq log\left(\frac{1}{P(X' \cup Y')}\right) - log\left(\frac{1}{P(X \cup Y)}\right) \iff \\
&log\left(\frac{P(X)}{P(X')}\right) \geq log\left(\frac{P(X \cup Y)}{P(X' \cup Y')}\right) \iff \frac{P(X)}{P(X')} \geq \frac{P(X \cup Y)}{P(X' \cup Y')} \iff \\
&\frac{P(X' \cup Y')}{P(X')} \geq \frac{P(X \cup Y)}{P(X)} \iff P(Y'|X') \geq P(Y|X) \iff \Delta P(Y|X) \geq 0 \ .
\end{aligned}$$

And in the same way: $\Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \equiv \Delta P(X|Y) \geq 0$. □

Proposition 4.3. *Satisfying SIM is a sufficient condition for satisfying the* IDENTITY, IDENTITY-SP, DEPENDENCY *and* UNEXPECTEDNESS *constraints.*

Proof A.2.4. The IDENTITY constraint states that $Sim(X,X) > Sim(X,XY)$. It is enough to prove that the object pairs (X,X) and (X,XY) satisfy the SIM conditions:

1.
$$\begin{aligned} PMI(X,X) - PMI(X,XY) = \\ = log\left(\frac{P(X,X)}{P(X)\cdot P(X)}\right) - log\left(\frac{P(X,XY)}{P(X)\cdot P(XY)}\right) = \\ = log\left(\frac{P(X)}{P(X)\cdot P(X)}\right) - log\left(\frac{P(XY)}{P(X)\cdot P(XY)}\right) = \\ = log\left(\frac{1}{P(X)}\right) - log\left(\frac{1}{P(X)}\right) = 0\ . \end{aligned}$$

2.
$$P(X \mid X) - P(XY \mid X) = 1 - \frac{P(XY,X)}{P(X)} = 1 - \frac{P(XY)}{P(X)} > 0\ .$$

3.
$$P(X \mid X) - P(X \mid XY) = 1 - 1 = 0\ .$$

Notice that the notation XY implies that X and Y are disjoint feature sets. Therefore, $P(XY) < P(X)$ and $\frac{P(XY)}{P(X)} < 1$. Therefore $1 - \frac{P(XY)}{P(X)} > 0$.

That is, the three conditions are satisfied and the second one is satisfied in an strict manner. Therefore, according to the SIM axiom: $Sim(X,X) > Sim(X,XY)$, satisfying the identity axiom. Notice that the second part of the IDENTITY constraint can be derived from its first part and the IDENTITY-SP constraint (see Section 4.3.2).

The IDENTITY-SP constraint states that $Sim(XY,XY) > Sim(X,X)$. It is enough to prove that (XY,XY) and (X,X) satisfies the SIM conditions:

1.
$$\begin{aligned} PMI(XY,XY) - PMI(X,X) = \\ = log\left(\frac{P(XY,XY)}{P(XY)\cdot P(XY)}\right) - log\left(\frac{P(X,X)}{P(X)\cdot P(X)}\right) = \\ = log\left(\frac{1}{P(XY)}\right) - log\left(\frac{1}{P(X)}\right) > 0\ . \end{aligned}$$

2.
$$P(XY \mid XY) - P(X \mid X) = 0\ .$$

3.
$$P(XY \mid XY) - P(X \mid X) = 0\ .$$

Therefore, the three conditions are satisfied and the first one is satisfied in an strict manner. Then, according to the SIM axiom: $Sim(XY,XY) > Sim(X,X)$, and the IDENTITY-SP constraint is complied.

To see that SIM captures the UNEXPECTEDNESS constraint, we need to prove that if $P(Y \mid X) > P(Y' \mid X)$ then $Sim(X,XY) > Sim(X,XY')$. It is enough to prove that (X,XY) and (X,XY') satisfies the SIM conditions expressed in Lemma 4.1. Notice that;

$$P(Y|X) > P(Y' \mid X) \Longrightarrow P(XY) > P(XY')\ .$$

Therefore:

1. $$\begin{aligned} PMI(X, XY) - PMI(X, XY') = \\ = log\left(\frac{P(X,XY)}{P(X)\cdot P(XY)}\right) - log\left(\frac{P(X,XY')}{P(X)\cdot P(XY')}\right) = \\ = log\left(\frac{P(XY)}{P(X)\cdot P(XY)}\right) - log\left(\frac{P(XY')}{P(X)\cdot P(XY')}\right) = \\ = log\left(\frac{1}{P(X)}\right) - log\left(\frac{1}{P(X)}\right) = 0\ . \end{aligned}$$

2. $$P(X \mid XY) - P(X \mid XY') = 0\ .$$

3. $$P(XY \mid X) - P(XY' \mid X) = P(Y \mid X) - P(Y' \mid X) > 0\ .$$

Therefore, the three conditions are satisfied and the second one is satisfied in an strict manner. Therefore, according to the SIM axiom: $Sim(XY, XY) > Sim(X, X)$, satisfying the UNEXPECTEDNESS constraint.

The dependency axiom states that if $P(XZ \mid YZ') > P(X \mid Y)$ and $P(YZ' \mid XZ) > P(Y \mid X)$ then $Sim(XZ, YZ') > Sim(X, Y)$. The three conditions of the SIM axiom for stating $Sim(XZ, YZ') > Sim(X, Y)$ are satisfied:

1. $$\begin{aligned} PMI(XZ, YZ') - PMI(X, Y) = \\ = log\left(\frac{P(XZ,YZ')}{P(XZ)\cdot P(YZ')}\right) - log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right) = \\ = log\left(\frac{P(XZ \mid YZ')}{P(XZ)}\right) - log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right) > \\ > log\left(\frac{P(X \mid Y)}{P(XZ)}\right) - log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right) > \\ > log\left(\frac{P(X \mid Y)}{P(X)}\right) - log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right) = \\ = log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right) - log\left(\frac{P(X,Y)}{P(X)\cdot P(Y)}\right) = 0\ . \end{aligned}$$

2. $$P(XZ \mid YZ') - P(X \mid Y) > P(X \mid Y) - P(X \mid Y) = 0\ .$$

3. $$P(YZ' \mid XZ) - P(Y \mid Z) > P(Y \mid X) - P(Y \mid X) = 0\ ,$$

where one inequality must be strict. Thus, SIM captures de DEPENDENCY constraint. □

Proposition 4.4. *SIM does not imply any constraint with respect to the ASYMMETRICITY conditions.*

Proof A.2.5. The Pointwise Mutual Information does not change in the assymmetricity condition:

$$PMI(XY, X) = PMI(X, XY) \ .$$

and the conditional probabilities grow in opposite directions:

$$P(XY \mid X) - P(X \mid XY) = -\big(P(X \mid XY) - P(XY \mid X)\big) \ .$$

Therefore, the SIM conditions never hold. □

Proposition 4.5:. *Under the assumption of statistical independence between the intersection and difference components of two fuzzy feature sets and considering the OIQ as the salience function:*

$$\mathcal{I}(X \cup Y) = \mathcal{I}(X \cap Y) + \mathcal{I}(X \setminus Y) + \mathcal{I}(Y \setminus X) \ .$$

SIM is equivalent to Tversky's Monotonicity axiom.

Proof A.2.6. The independence condition implies that:

$$\begin{aligned}
\mathcal{I}(X) &= \mathcal{I}(X \cap Y) + \mathcal{I}(X \setminus Y) \\
\mathcal{I}(Y) &= \mathcal{I}(X \cap Y) + \mathcal{I}(Y \setminus X) \\
\mathcal{I}(X \cup Y) &= \mathcal{I}(X \cap Y) + \mathcal{I}(X \setminus Y) + \mathcal{I}(Y \setminus X) \ .
\end{aligned}$$

Therefore:

$$\begin{aligned}
&\Delta\mathcal{I}(X) + \Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \\
&\quad \Longrightarrow 2\Delta\mathcal{I}(X \cap Y) + \Delta\mathcal{I}(X \setminus Y) + \Delta\mathcal{I}(Y \setminus X) \\
&\qquad \geq \Delta\mathcal{I}(X \cap Y) + \Delta\mathcal{I}(X \setminus Y) + \Delta\mathcal{I}(Y \setminus X) \\
&\quad \Longrightarrow \Delta\mathcal{I}(X \cap Y) \geq 0 \ ,
\end{aligned}$$

and

$$\begin{aligned}
&\Delta\mathcal{I}(X) \geq \Delta\mathcal{I}(X \cup Y) \\
&\quad \Longrightarrow \Delta\mathcal{I}(X \cap Y) + \Delta\mathcal{I}(X \setminus Y) \\
&\qquad \geq \Delta\mathcal{I}(X \cap Y) + \Delta\mathcal{I}(X \setminus Y) + \Delta\mathcal{I}(Y \setminus X) \\
&\quad \Longrightarrow 0 \geq \Delta\mathcal{I}(Y \setminus X) \ .
\end{aligned}$$

And in the same way:

$$\begin{aligned}
&\Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y) \\
&\quad \Longrightarrow \Delta\mathcal{I}(X \cap Y) + \Delta\mathcal{I}(Y \setminus X) \\
&\qquad \geq \Delta\mathcal{I}(X \cap Y) + \Delta\mathcal{I}(X \setminus Y) + \Delta\mathcal{I}(Y \setminus X) \\
&\quad \Longrightarrow 0 \geq \Delta\mathcal{I}(X \setminus Y) \ .
\end{aligned}$$

□

A.3. Formal Proofs for Chapter 5

Proposition 5.1. *Under the assumption that the PMI of two objects is not negative, the cosine similarity under the skip-gram with negative sampling representation satisfies* UNEXPECTEDNESS.

Proof A.3.1. First, the cosine similarity is a metric distance which satisfies maximality. Therefore, IDENTITY-SP can not be satisfied. However, according to several studies [74, 11], the scalar product of vectors in Skip-gram representation approaches PMI:

$$v_A \cdot v_B \simeq log\left(\frac{P(A,B)}{P(A)\cdot P(B)}\right) .$$

Therefore,

$$Cos(v_A, v_B) = \frac{v_A \cdot v_b}{\|v_A\| \cdot \|v_B\|} \simeq \frac{log\left(\frac{P(A,B)}{P(A)\cdot P(B)}\right)}{\left(log\left(\frac{1}{P(A)}\right)\cdot log\left(\frac{1}{P(A)}\right)\right)^{\frac{1}{2}}} .$$

In terms of Information Quantity, this can be expressed as:

$$Cos(v_A, v_B) \simeq \frac{\mathcal{I}(A) + \mathcal{I}(B) - \mathcal{I}(A,B)}{(\mathcal{I}(A)\cdot\mathcal{I}(B))^{\frac{1}{2}}} .$$

We can prove that UNEXPECTEDNESS is satisfied given that:

$$\begin{aligned} &P(Y \mid X) < P(Y' \mid X) \Longrightarrow P(XY \mid X) < P(XY' \mid X) \\ &\Longrightarrow \mathcal{I}(X) - \mathcal{I}(XY) < \mathcal{I}(X) - \mathcal{I}(XY') \Longrightarrow \mathcal{I}(XY) > \mathcal{I}(XY') . \end{aligned}$$

Therefore:

$$\begin{aligned} Cos(v_X, v_{XY}) &\simeq \frac{\mathcal{I}(X) + \mathcal{I}(XY) - \mathcal{I}(X, XY)}{(\mathcal{I}(X)\cdot\mathcal{I}(XY))^{\frac{1}{2}}} \\ &= \frac{\mathcal{I}(X)}{(\mathcal{I}(X)\cdot\mathcal{I}(XY))^{\frac{1}{2}}} < \frac{\mathcal{I}(X)}{(\mathcal{I}(X)\cdot\mathcal{I}(XY'))^{\frac{1}{2}}} = Cos(v_X, v_{XY'}) . \end{aligned}$$

□

Proposition 5.2. *Whenever $\alpha_1 = \alpha_4$, any variation in α_1 and α_1 in the ratio contrast model produces ordinal equivalent similarity functions.*

Proof A.3.2. Being $\alpha_1 = \alpha_4$, according to the ratio contrast model:

$$\begin{aligned} &Sim(X,Y) > Sim(X',Y') \Longleftrightarrow \\ &\frac{\alpha_1 \cdot f(X \cap Y)}{\alpha_2 \cdot f(X \setminus Y) + \alpha_3 \cdot f(Y \setminus X) + \alpha_1 \cdot f(X \cap Y)} > \\ &\frac{\alpha_1 \cdot f(X' \cap Y')}{\alpha_2 \cdot f(X' \setminus Y') + \alpha_3 \cdot f(Y' \setminus X') + \alpha_1 \cdot f(X' \cap Y')} . \end{aligned}$$

Knowing that the salience function f is positive, this is equivalent to:

$$\frac{1}{\frac{\alpha_2 \cdot f(X \setminus Y)}{\alpha_1 \cdot f(X \cap Y)} + \frac{\alpha_3 \cdot f(Y \setminus X)}{\alpha_1 \cdot f(X \cap Y)} + 1} > \frac{1}{\frac{\alpha_2 \cdot f(X' \setminus Y')}{\alpha_1 \cdot f(X' \cap Y')} + \frac{\alpha_3 \cdot f(Y' \setminus X')}{\alpha_1 \cdot f(X' \cap Y')} + 1} \iff$$
$$\frac{\alpha_2 \cdot f(X \setminus Y)}{\alpha_1 \cdot f(X \cap Y)} + \frac{\alpha_3 \cdot f(Y \setminus X)}{\alpha_1 \cdot f(X \cap Y)} < \frac{\alpha_2 \cdot f(X' \setminus Y')}{\alpha_1 \cdot f(X' \cap Y')} + \frac{\alpha_3 \cdot f(Y' \setminus X')}{\alpha_1 \cdot f(X' \cap Y')} \iff$$
$$\frac{\alpha_2 \cdot f(X \setminus Y) + \alpha_3 \cdot f(Y \setminus X)}{f(X \cap Y)} < \frac{\alpha_2 \cdot f(X' \setminus Y') + \alpha_3 \cdot f(Y' \setminus X')}{f(X' \cap Y')} .$$

which is not affected by α_1. □

The Ratio contrast model, does not capture the IDENTITY-SP *counter sample.*

Proof A.3.3. Assuming that the salience function f is zero for an empty set, the self similarity according to the ratio contrast model is:

$$Sim(X,X) = \frac{\alpha_1 \cdot f(X \cap X)}{\alpha_2 \cdot f(X \setminus X) + \alpha_3 \cdot f(X \setminus X) + \alpha_4 \cdot f(X \cap X)}$$
$$= \frac{\alpha_1 \cdot f(X)}{\alpha_2 \cdot f(\emptyset) + \alpha_3 \cdot f(\emptyset) + \alpha_4 \cdot f(X)} = \frac{\alpha_1 \cdot f(X)}{\alpha_4 \cdot f(X)} = \frac{\alpha_1}{\alpha_4} .$$

Therefore, the self similarity is fixed and it can be affected by the characteristics of the object, contradicting the IDENTITY-SP constraint. □

Proposition 5.3: *The PMI satisfies the* DEPENDENCY *constraint.*

Proof A.3.4. Under our probabilistic framework, if $P(XZ \mid YZ') > P(X \mid Y)$, then

$$PMI(XZ, YZ') = \log\left(\frac{P(XZ, YZ')}{P(XZ) \cdot P(YZ')}\right) = \log\left(P(XZ \mid YZ') \cdot \frac{1}{P(XZ)}\right) >$$
$$\log\left(P(X \mid Y) \cdot \frac{1}{P(XZ)}\right) = \log\left(\frac{P(X,Y)}{P(Y) \cdot P(XZ)}\right) > \log\left(\frac{P(X,Y)}{P(Y) \cdot P(X)}\right) = PMI(X,Y) .$$

□

Proposition 5.4: *The PMI does not satisfy* UNEXPECTEDNESS.

Proof A.3.5. When adding a feature set Z to an object X, the resulting PMI similarity to the original one is:

$$PMI(XZ, X) = \log\left(\frac{P(XZ, X)}{P(XZ) \cdot P(X)}\right) = \log\left(\frac{P(XZ)}{P(XZ) \cdot P(X)}\right)$$
$$= \log\left(\frac{1}{P(X)}\right) = \log\left(\frac{P(X)}{P(X) \cdot P(X)}\right) = PMI(X,X) .$$

Therefore PMI is constant and the UNEXPECTEDNESS constraint can not be satisfied. □

Proposition 5.5. *The ICM satisfies the SIM axiom when* $\alpha_1 + \alpha_2 > \beta > \alpha_1 > \alpha_2 > 0$.

Proof A.3.6. ICM is defined as:

$$\Delta ICM_{\alpha_1,\alpha_2,\beta}(X,Y) = \alpha_1 \cdot \Delta\mathcal{I}(X) + \alpha_2 \cdot \Delta\mathcal{I}(Y) - \beta \cdot \Delta\mathcal{I}(X \cup Y) .$$

Let us consider the case in which $\mathcal{I}(X \cup Y) \geq 0$. This proposition states that $\alpha_1 > 0$, $\alpha_2 > 0$. In addition, the SIM conditions state that $\Delta\mathcal{I}(X) \geq \Delta\mathcal{I}(X \cup Y)$ and $\Delta\mathcal{I}(Y) \geq$

$\Delta\mathcal{I}(X \cup Y)$. Therefore,

$$\begin{aligned}\Delta ICM_{\alpha_1,\alpha_2,\beta}(X,Y) &\geq \alpha_1 \cdot \Delta\mathcal{I}(X \cup Y) + \alpha_2 \cdot \Delta\mathcal{I}(X \cup Y) - \beta \cdot \Delta\mathcal{I}(X \cup Y) \\ &= (\alpha_1 + \alpha_2 - \beta) \cdot \Delta\mathcal{I}(X \cup Y) \ .\end{aligned}$$

Given that $\alpha_1 + \alpha_2 < \beta$, we can assert that:

$$(\alpha_1 + \alpha_2 - \beta) \cdot \Delta\mathcal{I}(X \cup Y) > 0 \ .$$

That is, the ICM increase is positive and the SIM axiom is satisfied.

Now, we consider the case $\mathcal{I}(X \cup Y) < 0$. The ICM conditions state that $\Delta\mathcal{I}(X) + \Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y)$. Then:

$$\begin{aligned}\Delta ICM_{\alpha_1,\alpha_2,\beta}(X,Y) &= \alpha_1 \cdot \Delta\mathcal{I}(X) + \alpha_2 \cdot \Delta\mathcal{I}(Y) - \beta \cdot \Delta\mathcal{I}(X \cup Y) \\ &\geq \alpha_1 \cdot \Delta\mathcal{I}(X) + \alpha_2 \cdot \Delta\mathcal{I}(Y) - \beta \cdot \Delta\mathcal{I}(X) - \beta \cdot \Delta\mathcal{I}(Y) \\ &= (\alpha_1 - \beta) \cdot \Delta\mathcal{I}(X) + (\alpha_2 - \beta) \cdot \Delta\mathcal{I}(Y) \\ &= -(\beta - \alpha_1) \cdot \Delta\mathcal{I}(X) - (\beta - \alpha_2) \cdot \Delta\mathcal{I}(Y) \ .\end{aligned}$$

Given that $\Delta\mathcal{I}(X) \geq \Delta\mathcal{I}(X \cup Y)$, $\Delta\mathcal{I}(Y) \geq \Delta\mathcal{I}(X \cup Y)$, $\beta > \alpha_1$ and $\beta > \alpha_2$, we can assert that:

$$\begin{aligned}&- (\beta - \alpha_1) \cdot \Delta\mathcal{I}(X) - (\beta - \alpha_2) \cdot \Delta\mathcal{I}(Y) \geq \\ \geq &- (\beta - \alpha_1) \cdot \Delta\mathcal{I}(X \cup Y) - (\beta - \alpha_2) \cdot \Delta\mathcal{I}(X \cup Y) = \\ = &- (2\beta - \alpha_1 - \alpha_2) \cdot \Delta\mathcal{I}(X \cup Y) \ .\end{aligned}$$

Given that $2\beta - \alpha_1 - \alpha_2 > 2\beta - \beta - \beta = 0$ and $\mathcal{I}(X \cup Y) < 0$:

$$-(2\beta - \alpha_1 - \alpha_2) \cdot \Delta\mathcal{I}(X \cup Y) > 0 \ .$$

□

Proposition 5.6. *The ICM generalizes pointwise mutual information and the product of conditional probabilities. Being* $\alpha_1 = \alpha_2 = 1 \leq \beta \leq 2$

$$\log(P(Occ(X) \mid Occ(Y)) \cdot P(Occ(Y) \mid Occ(X))) > ICM_{\alpha_1,\alpha_2,\beta}(X,Y) > PMI(Occ(X), Occ(Y)) \ .$$

Proof A.3.7. For the sake of simplicity, let us denote $P(Occ(X))$ as $P(X)$. When $\beta = \alpha_1 = \alpha_2 = 1$ then ICM matches with the Pointwise Mutual Information.

$$ICM_{1,1,1}(X,Y) = log\left(\frac{P(X,Y)^{\beta}}{P(X)^{\alpha_1} \cdot P(Y)^{\alpha_2}}\right) = log\left(\frac{P(X,Y)}{P(X) \cdot P(Y)}\right) \ .$$

At the opposite extrem when $\beta = \alpha_1 + \alpha_2 = 2$, the ICM fit into the product of conditional

probabilities.

$$ICM_{1,1,2}(X,Y) = log\left(\frac{P(X,Y)^2}{P(X)^1 \cdot P(Y)^1}\right) =$$

$$= log\left(\frac{P(X,Y)\cdot P(X,Y)}{P(X)\cdot P(Y)}\right) = log\big(P(Y \mid X)\cdot P(X \mid Y)\big)\ .$$

□

Proposition 5.7. *If independence is assumed between the features in the intersection and difference subsets in the ICM and the information quantity is used as the salience function in the linear contrast model, then the ICM and the linear contrast model are equivalent.*

$$ICM_{\alpha_1,\alpha_2,\beta}(X,Y) = (\alpha_1+\alpha_2-\beta)\cdot\mathcal{I}(X\cap Y)-(\beta-\alpha_1)\cdot(\mathcal{I}(X\setminus Y))-(\beta-\alpha_2)\cdot(\mathcal{I}(Y\setminus X))\ .$$

Proof A.3.8.

$$\begin{aligned} &ICM_{\alpha_1,\alpha_2,\beta}(X,Y) = \alpha_1\cdot\mathcal{I}(X)+\alpha_2\cdot\mathcal{I}(Y)-\beta\cdot\mathcal{I}(X\cup Y)\\ &= \alpha_1\cdot(\mathcal{I}(X\cap Y)+\mathcal{I}(X\setminus Y))+\alpha_2\cdot(\mathcal{I}(X\cap Y)+\mathcal{I}(X\setminus Y)))-\\ &\qquad \beta\cdot(\mathcal{I}(X\cap Y)+\mathcal{I}(X\setminus Y)+\mathcal{I}(Y\setminus X))\\ &= (\alpha_1+\alpha_2-\beta)\cdot\mathcal{I}(X\cap Y)-(\beta-\alpha_1)\cdot\mathcal{I}(X\setminus Y)-(\beta-\alpha_2)\cdot\mathcal{I}(Y\setminus X)\ . \end{aligned}$$

□

A.4. Formal Proofs for Chapter 7

Proposition 7.1. *The OIQ-based ranking fusion approach satisfies monotonicity, dependence and cancellation.*

Proof A.4.1. Let's prove that OIQ satisfies monotonicity. Le be Γ the feature set $\{\gamma_1,..,\gamma_n\}$ and let be γ another feature and $\gamma_{d_1\leftrightarrow d_2}$ the result of swapping the scores of d_1 and d_2 in γ like in the Property 7.1.

According to the definition of the OIQ based ranking fusion, $F_{\Gamma\cup\{\gamma\}}(d_1)$ and $F_{\Gamma\cup\{\gamma_{d_1\leftrightarrow d_2}\}}(d_1)$ correspond with the information quantity of the following observation outcomes:

$$\mathcal{O}_{\Gamma\cup\{\gamma\}}(d_1) = (\gamma_1(d_1),..,\gamma_n(d_1),\gamma(d_1))$$
$$\mathcal{O}_{\Gamma\cup\{\gamma_{d_1\leftrightarrow d_2}\}}(d_1) = (\gamma_1(d_1),..,\gamma_n(d_1),\gamma(d_2))$$

According to Property 3.1 (Feature Value Monotonicity), given that $\gamma(d_1) > \gamma(d_2)$, then:

$$\mathcal{I}_{\Gamma\cup\{\gamma\}}(d_1) > \mathcal{I}_{\Gamma\cup\{\gamma_{d_1\leftrightarrow d_2}\}}(d_1)$$

Therefore, the Strict Monotonicity property of ranking fusion is satisfied. The dependency and cancellation properties are also satisfied. They are directly connected with properties 3.5 and 3.6.

Proposition 7.2. *The UIR has the following correspondence with the OIQ:*

$$UIR(d) \simeq 2^{I_{\{\gamma_1,\ldots,\gamma_n\}}(d)} - 2^{I_{\{-\gamma_1,\ldots,-\gamma_n\}}(d)} \ .$$

Proof A.4.2. From the definition of the OIQ approach, we have:

$$2^{\mathcal{I}_\Gamma\left(d\right)} = P\big(\{d' \in \mathcal{D} : rank_\gamma(d') \leq rank_\gamma(d), \forall\ \gamma \in \Gamma\}\big) \ .$$

Assume than the set of features is $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$, if we consider the set with the opposite features, we have that,

$$2^{\mathcal{I}_{\{-\gamma_1,\ldots,-\gamma_n\}}\left(d\right)} = P\big(\{d' \in \mathcal{D} : rank_{-\gamma}(d') \geq rank_{-\gamma}(d), \forall\ \gamma \in \Gamma\}\big) \ .$$

Therefore, it is possible to write UIR easily as:

$$UIR(d) \simeq 2^{I_{\{\gamma_1,\ldots,\gamma_n\}}(d)} - 2^{I_{\{-\gamma_1,\ldots,-\gamma_n\}}(d)} \ .$$

□

Bibliography

[1] Abhijit Adhikari, Shivang Singh, Deepjyoti Mondal, Biswanath Dutta, and Animesh Dutta. A novel information theoretic framework for finding semantic similarity in wordnet. *CoRR*, abs/1607.05422, 2016.

[2] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[3] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45 – 65, 2003.

[4] Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In *Proceedings of Machine Translation Summit VIII*, pages 15–20, 2001.

[5] Joshua Albrecht and Rebecca Hwa. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 187–190, 2008.

[6] Enrique Amigó, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of replab 2014: Author profiling and reputation dimensions for online reputation management. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings*, 09/2014 2014.

[7] Enrique Amigó, Fernando Giner, Julio Gonzalo, and Felisa Verdejo. On the foundations of similarity in information access. *Inf. Retr. J.*, 23(3):216–254, 2020.

[8] Enrique Amigó, Fernando Giner, Stefano Mizzaro, and Damiano Spina. A formal account on effectiveness evaluation and ranking fusion. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2018, Tianjin, China September 14-17, 2018*, 2018.

[9] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. *J. Artif. Intell. Res. (JAIR)*, 42:689–718, 2011.

[10] Andrea Argentini. *Ranking aggregation based on belief function theory*. PhD thesis, University of Trento, 2012.

[11] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *TACL*, 4:385–399, 2016.

[12] Javier Artiles, Enrique Amigó, and Julio Gonzalo. The Role of Named Entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, EMNLP '09, pages 534–542. Association for Computational Linguistics, 2009.

[13] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. The SemEval-2007 WePS evaluation: Establishing a Benchmark for the Web People Search Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 64–69, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[14] E. G. Ashby and N. A. Perrin. Toward a unified theory of similarity and recognition. *Psychological review*, 95(1):124–150, 1988.

[15] Javed A Aslam and Mark Montague. Models for metasearch. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–284, 2001.

[16] Fred Attneave. Dimensions of similarity. *American Journal of Psychology*, 63(4):516–556, 1950.

[17] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.

[18] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[19] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[20] Joseph P. Bullinaria, John A.and Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, 2007.

[21] Luca Cazzanti and Maya Gupta. Information-theoretic and Set-theoretic Similarity. In *2006 IEEE International Symposium on Information Theory*, pages 1836–1840. IEEE, July 2006.

[22] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions, 2007.

[23] Kenneth W. Church and William A. Gale. Poisson mixtures. *Natural Language Engineering*, 1:163–190, 1995.

[24] G. V. Cormack, C. L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods.

[25] Simon Corston-Oliver, Michael Gamon, and Chris Brockett. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 140–147, 2001.

[26] Valerie Cross. Fuzzy information retrieval. *Journal of Intelligent Information Systems*, 3(1):29–56, Feb 1994.

[27] Ronan Cummins and Colm O'Riordan. Analysing ranking functions in information retrieval using constraints. 2009.

[28] Ido Dagan, Fernando Pereira, and Lillian Lee. Similarity-based estimation of word cooccurrence probabilities. In *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 272–278, 1994.

[29] Hoa Trang Dang. Overview of DUC 2005. In *Proceedings of the 2005 Document Understanding Workshop*, 2005.

[30] Hoa Trang Dang. Overview of DUC 2006. In *Proceedings of the 2006 Document Understanding Workshop*, 2006.

[31] Jean C. de Borda. *Memoire sur les Elections au Scrutin.* Histoire de l'Academie Royale des Sciences, Paris, 1781.

[32] M. de Condorcet. *Essai Sur l'Application de l'Analyse À la Probabilite des Decisions Rendues e la Pluralite des Voix.* 1785.

[33] Aldo De Luca and Settimo Termini. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and control*, 20(4):301–312, 1972.

[34] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

[35] M Delgado, MJ Martín-Bautista, D Sánchez, and MA Vila. Aggregating opinions in an information retrieval problem. In *Proc. of EUROFUSE Workshop on Preference Modelling and Applications, Granada, Spain*, pages 169–173, 2001.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[38] Wei Dong, Moses Charikar, and Kai Li. Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 123–130. ACM, 2008.

[39] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

[40] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160, 2003.

[41] Hui Fang, Tao Tao, and Chengxiang Zhai. Diagnostic evaluation of information retrieval models. *ACM Trans. Inf. Syst.*, 29(2):7:1–7:42, April 2011.

[42] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 480–487, New York, NY, USA, 2005. ACM.

[43] Hui Fang and ChengXiang Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122. ACM, 2006.

[44] Norbert Fuhr, Marc Lechtenfeld, Benno Stein, and Tim Gollub. The optimum clustering framework: Implementing the cluster hypothesis. *Information Retrieval*, 15:93–115, 2012.

[45] A. Garg, C. G. Enright, and M. G. Madden. On asymmetric similarity search. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 649–654, Dec 2015.

[46] Jean Mark Gawron. Improving sparse word similarity models with asymmetric measures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 296–301, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[47] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86, 2010.

[48] Fernando Giner, Enrique Amigó, and Felisa Verdejo. Integrating learned and explicit document features for reputation monitoring in social media. *Knowledge and Information Systems*, 62(3):951–985, 2020.

[49] R. L. Goldstone, D. L. Medin, and D. Gentner. Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, 23:222–264, 1991.

[50] Sreenivas Gollapudi and Rina Panigrahy. Exploiting asymmetry in hierarchical topic extraction. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 475–482, New York, NY, USA, 2006. ACM.

[51] Jacinto González-Pachón and Carlos Romero. Aggregation of partial ordinal rankings: an interval goal programming approach. *Computers & Operations Research*, 28(8):827–834, 2001.

[52] A. Gordo, F. Perronnin, Y. Gong, and S. Lazebnik. Asymmetric distances for binary embeddings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):33–47, 2014.

[53] Warren R Greiff and Jay M Ponte. The maximum entropy approach and probabilistic ir models. *ACM Transactions on Information Systems (TOIS)*, 18(3):246–287, 2000.

[54] S. P. Harter. A probabilistic approach to automatic keyword indexing. part ii: An algorithm for probabilistc indexing. *Journal of the American Society for Informaiton Science*, 26(4):280–289, 1975.

[55] F Herrera, E Herrera-Viedma, and Luis Martínez. An information retrieval system with unbalanced linguistic information based on the linguistic 2-tuple model. In *8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Bases Systems (IPMU'2002). Annecy (France)*, pages 23–29, 2002.

[56] W. E. Hick. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4:11–26, 1952.

[57] D. Hiemstra and W. Kraaij. Twenty-one at trec-7: ad-hoc and cross-language track. In E.M. Voorhees and D.K. Harman, editors, *Seventh Text REtrieval Conference, TREC 1998*, volume 500-24 of *NIST Special Publications*, pages 227–238, Gaithersburg, MD, USA, 1998. National Institute of Standards and Technology (NIST).

[58] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[59] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[60] Mihir Jain, Hervé Jégou, and Patrick Gros. Asymmetric hamming embedding: Taking the best of our bits for large scale image search. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, page 1441–1444, New York, NY, USA, 2011. Association for Computing Machinery.

[61] Yang Jiao, Matthieu Cornec, and Jérémie Jakubowicz. An entropy-based term weighting scheme and its application in e-commerce search engines. In *International Symposium on Web Algorithms*, 2015.

[62] Harry Joe. *Multivariate models and multivariate dependence concepts*. CRC Press, 1997.

[63] Arnold Kaufmann. *Introduction to the theory of fuzzy subsets*, volume 2. Academic Pr, 1975.

[64] Weimao Ke. Information-theoretic term weighting schemes for document clustering. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 143–152. ACM, 2013.

[65] J. Kohlas, M. Pouly, and C. Schneuwly. Information algebra. In Benjamin Wah, editor, *Wiley Encyclopedia of Computer Science and Engineering*. John Wiley & Sons, Inc., 2008.

[66] Jürg Kohlas. Algebras of information. a new and extended axiomatic foundation. *arXiv preprint arXiv:1701.02658*, 2017.

[67] Bart Kosko. Fuzziness vs. probability. *International Journal of General System*, 17(2-3):211–240, 1990.

[68] C. Krumhansl. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, 85(5):445–463, 1978.

[69] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[70] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, May 2003.

[71] LDC. Linguistic Data Annotation Specification: Assessment of Adequacy and Fluency in Translations. Revision 1.5. Technical report, Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess04.pdf, 2005.

[72] JH Lee. Analysis of multiple evidence combination. 20th acm sigir conf. on research and development in information retrieval (sigir 97). *New York*, pages 267–276, 1997.

[73] Sungjin Lee, Gyunyoung Heo, and Soon Heung Chang. Prediction of the human response time with the similarity and quantity of information. *Reliability Engineering & System Safety*, 91(6):728 – 734, 2006.

[74] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *journal*, pages 2177–2185, 2014.

[75] Chin-Yew Lin. Rouge: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[76] Dekang Lin. Dependency-based evaluation of minipar. In *Proc. Workshop on the Evaluation of Parsing Systems*, Granada, 1998.

[77] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[78] Maurice HT Ling. Distance coefficients between two lists or sets, 2010.

[79] Ding Liu and Daniel Gildea. Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of the 2007 Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 41–48, 2007.

[80] Kenneth O May. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica: Journal of the Econometric Society*, pages 680–684, 1952.

[81] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

[82] Doug Medin, Robert Goldstone, and Dedre Gentner. Respects for similarity. 100:254–278, 04 1993.

[83] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*, Boston, Massachusetts, July 2006.

[84] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[85] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[86] David R. H. Miller, Tim Leek, and Richard M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 214–221, New York, NY, USA, 1999. ACM.

[87] Mark H. Montague and Javed A. Aslam. Condorcet fusion for improved retrieval. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 4-9, 2002*, pages 538–548. ACM, 2002.

[88] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[89] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, 2000.

[90] Rabia Nuray-Turan and Fazli Can. Automatic ranking of retrieval systems using fusion data. 42:595–614, 05 2006.

[91] A. Oktoveri, A. T. Wibowo, and A. M. Barmawi. Non-relevant document reduction in anti-plagiarism using asymmetric similarity and avl tree index. In *2014 5th International Conference on Intelligent and Advanced Systems (ICIAS)*, pages 1–5, June 2014.

[92] Andrew Ortony, Richard J Vondruska, Mark A Foss, and Lawrence E Jones. Salience, similes, and the asymmetry of similarity. *Journal of Memory and Language*, 24(5):569 – 594, 1985.

[93] Kishore Papineni. Why inverse document frequency? In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[94] D. Partridge and W. Krzanowski. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology*, 39(10):707 – 717, 1997.

[95] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[96] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[97] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[98] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[99] R.I. Why RI? business through data-driven reputation management, 2018.

[100] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In *Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, SIGIR '80, pages 35–56, Kent, UK, UK, 1981. Butterworth & Co.

[101] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[102] E. Rosch. Cognitive reference points. *Cognitive Psychology*, 7:532–547, 1975.

[103] E. Z. Rothkopf. A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, 53:94–101, 1957.

[104] Imre J Rudas and M Okyay Kaynak. Entropy-based operations on fuzzy sets. *IEEE transactions on fuzzy systems*, 6(1):33–40, 1998.

[105] R.N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323, 1987.

[106] Roger N Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140, 1962.

[107] Yangyang Shi, Pascal Wiggers, and Catholijn M. Jonker. Towards recurrent neural networks language models with linguistic and contextual features. pages 1664–1667. ISCA, 2012.

[108] Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. Idf for word n-grams. *ACM Trans. Inf. Syst.*, 36(1):5:1–5:38, June 2017.

[109] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 66–73, New York, NY, USA, 2001. ACM.

[110] Anselm Spoerri. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. 43(4):1059–1070, 2007.

[111] Joshua B Tenenbaum and Thomas L Griffiths. Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, 24(4):629–640, 2001.

[112] Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*, 1997.

[113] Antonio Toral, Pavel Pecina, Longyue Wang, and Josef van Genabith. Linguistically-augmented perplexity-based data selection for language models. *Computer Speech and Language*, 32(1):11 – 26, 2015. Hybrid Machine Translation: integration of linguistics and statistics.

[114] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[115] Amos Tversky and Itamar Gati. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123, 1982.

[116] Christopher C. Vogt and Garrison W. Cottrell. Predicting the performance of linearly combined ir systems. pages 190–196. ACM, 1998.

[117] Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, NY, USA, 2006. ACM.

[118] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[119] Shengli Wu, Chunlan Huang, Liang Li, and Fabio Crestani. Fusion-based methods for result diversification in web search. *Information Fusion*, 45:16 – 26, 2019.

[120] Ronald R Yager and Alexander Rybalov. On the fusion of documents from multiple collection information retrieval systems. *Journal of the American Society for Information Science*, 49(13):1177–1184, 1998.

[121] Ning Yu. A one-shot proof of Arrow's impossibility theorem. *Economic Theory*, 50(2):523–525, June 2012.

[122] Lotfi Asker Zadeh. Probability measures of fuzzy events. *Journal of mathematical analysis and applications*, 23(2):421–427, 1968.

[123] ChengXiang Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2(3):137–213, March 2008.

[124] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.