Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 153-156 recibido 18-12-2021 revisado 01-02-2022 aceptado 09-02-2022

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

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

Fernando Giner Martínez
Research Group in Natural Language Processing and Information Retrieval
Universidad Nacional de Educación a Distancia (UNED)
C/ Juan del Rosal, 16, 28040 - Madrid, Spain
fginer3@gmail.com

Abstract: This is a summary of the Ph.D. thesis written by Fernando Giner Martínez at National Distance University ETSI - UNED, under the supervision of Ph.D. Enrique Amigó Cabrera. The author was examined on Thursday, September 23th, 2021 by a committee formed by Ph.D. Fermín Moscoso del Prado Martín from Lingvist and Radbound University Nijmegen, Ph.D. Julián Urbano from University of Delf and Ph.D. Victor Fresno Fernández from UNED. The Ph.D. thesis obtained Summa Cum Laude.
Keywords: Document representation, information theory, similarity, ranking fusion.

Resumen: Este es un resumen de la tesis doctoral realizada por Fernando Giner Martínez en la Universidad Nacional de Educación a Distancia ETSI - UNED bajo la dirección del doctor D. Enrique Amigó Cabrera. El acto de defensa tuvo lugar el jueves 23 de septiembre de 2021 ante el tribunal formado por los doctores D. Fermín Moscoso del Prado Martín de Lingvist y de Radbound University Nijmegen, D. Julián Urbano de la Universidad de Delf y D. Victor Fresno Fernández de la UNED. La tesis obtuvo la calificacion de Sobresaliente Cum Laude por unanimidad.
Palabras clave: Representación de documentos, teoria de la información, similaridad, fusión de ránkings.

1 Introduction

Information Access is a research area which involves many tasks such as Text Mining, Information Retrieval or Text Categorization. In all these tasks, document representation is a key step. Document features can be binary, such as word occurrence, named entities, links, or any kind of linguistic structure. Other features are defined in a continuous range, such as time stamp, topicality, sentiment polarity, etc.

We can highlight three main issues in document representation. First, features have a certain 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, "Obama" does not contribute substantially to the information provided by "Barak Obama", given that "Obama" is informative enough in this context. The third issue is feature scaling. For instance, time stamps and word occurrences are completely different scales.

These three issues are tackled in a different way depending on whether we are in a supervised or unsupervised scenario. In the first 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 scale properties of features. For instance, a supervised classifier learns that "Obama" is more relevant than "said" when classifying news by topic. It can also infer that "Barak" does not provide additional evidence regarding "Obama", and also that news published less than 72 hours ago are more relevant for readers. Although in some contexts supervised approaches have shown to be highly effective, their drawbacks have been widely discussed in the literature, such as overfitting, domain dependency, data bias, annotation cost, etc. Another important drawback is that supervised learning does not provide mechanisms to manage information pieces, e.g., aggregation or comparison operators.

On the other hand, in the absence of human annotated data, the weight, dependence or scale of features is determined according to their distribution in a document collection. Typically, unexpected features have more presence in the representation than expected feature values. For instance, the word-feature "Obama" has more weight in the representation than frequent common words. The feature dependency can be also inferred from coocurrence. For instance, "Barak" and "Obama" are two word features which tends to appear together. As discussed in the first chapters of the thesis, the unexpectedness and coocurrence of features is the basis of the popular tf.idf feature weighting, stopwords removal or 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, we can estimate the probability of a word, n-gram, tag, etc. However, the likelihood of values in continuous features, for instance time stamp, depends on the granularity in which time is discretized (i.e. days, minutes, etc.). As far as we know, there are not standard criteria to quantify the likelihood of continuous feature values in the context of document representation for Information Access.

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

ORF has relevant implications in different lines. In this thesis we delved into three of them. First, it provides a common theoretical framework to analyse, compare and generalise document similarity functions which are based on different representation schemes. Second, it allows to integrate intrinsic and extrinsic document features in the same representation. Intrinsic features includes words, n-grams, etc. Extrinsic features can be the output of a clustering process or category membership values generated by classification systems. And third, it provides us a theoretical foundation and mechanism for ranking fusion.

2 General outline of the dissertation

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

Chapter 1 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 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 Our representation framework ORF is presented (Giner, Amigó, and Verdejo, 2020; Giner and Amigó, 2016). It 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 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 (Amigó et al., 2020; Amigó et al., 2017a).

Chapter 5 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) (Amigó et al., 2011). 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 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 (Giner, Amigó, and Verdejo, 2020).

Chapter 7 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 (Amigó et al., 2017b; Amigó et al., 2018).

Chapter 8 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.

3 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 formalising a number of desirable properties, we observe that none of the families of document representation frameworks (e.g. set-based, metric spaces, language models, etc.) complies with all constraints.

On the basis of this analysis, the second and main contribution in this thesis is the definition of the Observational Representation Framework (ORF), which extends the traditional Shannon's notion of Information Content ($-log(P(x))$) to the management of continuous feature values. This is called the Observational Information Quantity (OIQ) and is grounded on feature fuzzy sets and inclusion relationships between document observation outcomes in a document collection. We study in a comprehensive way the for-

mal properties of ORF and OIQ as well as their generalization power regarding traditional representation approaches.

The third contribution is the analysis of similarity functions and their foundations (i.e. cosine, euclidean, feature overlap, etc.). We will see, through the study of counterexamples and evidence provided in the literature, that euclidean axioms, as well as set-based axioms (Tversky's model) do not capture similarity properly in the context of information access systems. On the basis of ORF, we review the axiomatic in 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 parametrisable similarity function called Information Contrast Model (ICM). ICM, besides satisfying desirable formal constraints, it generalises traditional functions such as PMI, conditional probability, euclidean distance or Tversky's Linear Contrast Model.

The fourth contribution is related with the capability of ORF to aggregate heterogenous features in a document representation. For this, we develop a study case: clustering of tweets in the context of on-line reputation management. We prove empirically that the model integrates effectively discrete features (words) with continuous feature values. In our study case, continuous values are the proximity to pre-annotated categories of tweets and previously generated clusters. The results show that adding heterogeneous features increases the similarity predictive power between tweet representations. In this sense, 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 fusion ranking fusion on the basis of OIQ and ORF. The application of our framework in 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) generalises traditional ranking fusion algorithms and explains the effectiveness of existing approaches under different situations. We study empirically these phenomena on six different ranking fusion scenarios.

Acknowledgements

The work was supported by the Ministerio de Economía y Competitividad, TIN Program (Vemodalen), under Grant Number: TIN2015−71785−R.

References

Amigó, E., F. Giner, J. Gonzalo, and F. Verdejo. 2017a. An axiomatic account of similarity. In *Proceedings of the SIGIR'17 Workshop on Axiomatic Thinking for Information Retrieval and Related Tasks (ATIR)*, SIGIR '20, New York, NY, USA. ACM.

Amigó, E., F. Giner, J. Gonzalo, and F. Verdejo, 2017b. *A Formal and Empirical Study of Unsupervised Signal Combination for Textual Similarity Tasks*, pages 369–382. Springer International Publishing, Cham.

Amigó, E., F. Giner, J. Gonzalo, and F. Verdejo. 2020. On the foundations of similarity in information access. *Inf. Retr. J.*, 23(3):216–254.

Amigó, E., F. Giner, S. Mizzaro, and D. Spina. 2018. 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.

Amigó, E., J. Gonzalo, J. Artiles, and F. Verdejo. 2011. Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks. *J. Artif. Intell. Res. (JAIR)*, 42:689–718.

Giner, F. and E. Amigó. 2016. General representation model for text similarity. In *Proceedings of FETLT'16*.

Giner, F., E. Amigó, and F. Verdejo. 2020. Integrating learned and explicit document features for reputation monitoring in social media. *Knowledge and Information Systems*, 62(3):951–985.