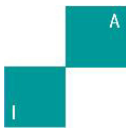
Inteligencia Artificial, 22(63) (2019), 61-80
doi: 10.4114/intartif.vol22iss63pp61-80



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

A Flexible Supervised Term-Weighting Technique and its Application to Variable Extraction and Information Retrieval

Mariano Maisonnave[1], Fernando Delbianco[2], Fernando Tohmé[2], Ana Maguitman[1]

[1] Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur,
Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET),
Bahía Blanca, Argentina
{mariano.maisonnave,agm}@cs.uns.edu.ar

[2] Departamento de Economía, Universidad Nacional del Sur,
Instituto de Matemática de Bahía Blanca (UNS-CONICET),
Bahía Blanca, Argentina
fernando.delbianco@uns.edu.ar, ftohme@criba.edu.ar

Abstract

Successful modeling and prediction depend on effective methods for the extraction of domain-relevant variables. This paper proposes a methodology for identifying domain-specific terms. The proposed methodology relies on a collection of documents labeled as relevant or irrelevant to the domain under analysis. Based on the labeled document collection, we propose a supervised technique that weights terms based on their descriptive and discriminating power. Finally, the descriptive and discriminating values are combined into a general measure that, through the use of an adjustable parameter, allows to independently favor different aspects of retrieval such as maximizing precision or recall, or achieving a balance between both of them. The proposed technique is applied to the economic domain and is empirically evaluated through a human-subject experiment involving experts and non-experts in Economy. It is also evaluated as a term-weighting technique for query-term selection showing promising results. We finally illustrate the applicability of the proposed technique to address diverse problems such as building prediction models, supporting knowledge modeling, and achieving total recall.

Keywords:Term Weighting, Variable Extraction, Information Retrieval, Query-Term Selection

1 Introduction

A great number of machine learning and data science applications require identifying domain- or topic-relevant terms. For instance, automatic query formulation requires selecting good query terms; classification requires extracting good features, and in general, any modeling and prediction task requires mechanisms for variable extraction as an initial step to build useful representations. Also, term weighting is a crucial component of these representations since the importance of a term for a domain or topic can usually be numerically estimated and such weights have an impact on the task to be carried out.

Several term-weighting schemes have been proposed in the literature with varying degree of success. Most of these methods apply an unsupervised approach to determine term importance. This is the case of the widely-used TF-IDF weighting scheme, where terms are weighted based on local (TF) and global (IDF) term frequencies, but no class-label information is used to compute these weights. This scheme is limited when it comes to identifying terms that are important for general topics or domains because it has the constraints of being document dependent

ISSN: 1137-3601 (print), 1988-3064 (on-line)

(as it is based on the document and not on the general topic or domain) and label independent (as it is independent of the topic or domain label). Other term weighting methods take a supervised approach to assess the importance of a term in a class. However, term importance is typically taken as a fixed value independent of the task at hand. This represents a limitation because the importance of a term depends on whether the term is needed for query construction, clustering, classification, document summarization, among other tasks. Even for a specific task, such as is the case of query construction, a term may be more or less effective depending on whether the application requires high recall (e.g., looking for all relevant literature about a given topic) or high precision (e.g., looking for a specific piece of information such as a date, place or name). For example, a term that is a useful descriptor for a topic of interest, and therefore useful for attaining high recall, may lack discriminating power, resulting in low precision, unless it is combined with other terms that can discriminate between good and bad results.

This paper proposes a methodology that can be applied to identify domain- or topic-relevant variables from labeled documents. Two forms of relevance are distinguished, namely the relevance of a term as a descriptor, or *descriptive relevance*, and the relevance of a term as a discriminator, or *discriminative relevance*. Guided by this distinction, we propose two weighting schemes that account for these two notions of relevance. These weights are then combined into a parameter-dependent measure to which we refer to as FDD_β, accounting for a general notion of relevance. As we will show in the experiments, the FDD_β measure offers an advantage over several state-of-the-art term-weighting schemes as its parameter can be adjusted to emphasize different aspects of relevance (i.e., descriptive and discriminative relevance). As a consequence, the FDD_β measure has the practical implication of being able to favor either precision or recall, as well as to achieve a balance between both.

The paper is structured as follows. Section 2 briefly describes background concepts and reviews existing term-weighting schemes. Section 3 presents our novel term-weighting scheme, to which we refer to as FDD_β. Section 4 describes the data collection used in our analysis and evaluates FDD_β through a user study and as a query-term selection mechanisms. Section 5.1 illustrates the application of the proposal in variable extraction, knowledge modeling and information retrieval. Finally, section 6 presents the conclusions and outlines future research work.

2 Background and Related Work

Term weighting has been widely used in text classification and information retrieval. For historical reasons, term-weighting methods in text classification were originally borrowed from the information retrieval area, which traditionally applied unsupervised techniques. These traditional term-weighting schemes were designed to improve both recall and precision in the retrieval task. Based on these considerations, Salton and Buckley (1988) claimed that at least three main factors are required in any term weighting scheme. The first is a local factor that stands for the presence of the term in the document. This factor represents whether the term appears at all, and how many times it does. It represents the idea that frequent terms are semantically close to the content of the document. Such a factor is designed to improve recall. The second factor is a global value associated with each term, which represents how frequent the term is in the document collection, in such a way that frequent terms are penalized. The rationale for using this penalizing factor is that common terms are poor discriminators, and as a consequence, they are not useful to tell apart among different documents containing them. It is known that using this factor helps to achieve higher precision. Note, however, that this might be at the expenses of a drop in recall. Finally, the terms are sometimes corrected by a normalization factor.

The simplest local factor is the binary one, which only measures the presence or absence of the term in the document (with values 1 or 0). Another simple and highly-used factor is *term frequency* (TF), which counts the number of times a term appears in a document. It relies on the assumption that most frequent terms are closely related to the content of the document. Leopold and Kindermann (2002) propose *inverse term frequency* (ITF) as an alternative to the classic TF. The ITF weight is based on Zipf's Law and normalizes the local factor to the interval [0,1]. On the other hand, Debole and Sebastiani (2004) propose another variation for the local factor, with a logarithmic transformation in which the terms that are extremely frequent do not increase at the same rate as in TF. Hassan et al. (2007) present a new local factor using a variant of *TextRank* [23] as a scoring function, which recursively increases the importance of a term by determining the degree of connectivity between other terms using co-occurrence as a way to measure connectivity. *TextRank* is based on the renowned *PageRank* algorithm [25].

A simple global factor can be computed by counting the number of documents in the corpus that contain the term. We refer to this factor as *term global frequency* (TGF). The best known global factor is the *inverse document frequency* (IDF) function [28], which relies on the assumption that terms that occur in many documents are not good for discrimination. The TF-IDF formulation is a widely used weighting scheme because it reaches a good balance between the local (TF) and the global (IDF) factor. Tokunaga and Makoto (1994) propose a variant of IDF named *weighted inverse document frequency* (WIDF) that penalizes frequent terms by taking into

account the number of times they occur in each document of a collection. A variant of TF-IDF called *modified inverse document frequency* (MIDF) that combines TF and WIDF is proposed in [4]. According to the authors, MIDF outperforms TF-IDF in text classification. Also, they remark the ability of MIDF to adapt to dynamic document corpora.

While unsupervised weighting schemes have proved to be useful in many scenarios, these methods do not take full advantage of class information, which is available as part of the training set in a class-labeled collection. The design of term-weighting methods that exploit class information gained increasing attention, giving rise to different forms of supervised term-weighting schemes [2, 3, 6, 8, 9, 14, 31, 32]. A simple method that uses class information can be computed by counting the number of documents in a class that contain the term. We use TGF* to refer to this method. Another supervised weighting scheme is the *inverse class frequency factor* (ICF), which relies on the assumption that a term that occurs in documents from a single class are good discriminants of that class. Conversely, terms that appear in documents from different classes contribute poorly to the identification of the class of the documents. So, this factor penalizes a term proportionally to the number of different classes in which the term appears. Other functions from traditional information theory such as *mutual information* (MI), *chi-squared* (χ^2), *information gain* (IG) and *gain ratio* (GR) can be used as supervised term-weighting scores to capture the idea that the most valuable terms for categorization under a class are those that are distributed most differently in the sets of positive and negative examples of the class. A classic feature scoring function that is commonly used as a global term-weighting factor is the *odds ratio* (OR) [30]. This score is based on the conditional probability of a term occurring given a class. Another supervised technique known as *category relevance factor* (CRF) computes a factor that stands for the discriminating power of a feature to a class [5]. Some feature selection techniques that were adapted for term weighing are the *Galavotti-Sebastiani-Simi* coefficient (GSS) [11] and *entropy-based category coverage difference* (ECCD) [16]. Liu et al. (2009) propose a probabilistic-based technique (Prob) that involves two ratios directly related to the term's strength in representing a category. These ratios are such that one of them increases if the term appears in a lot of documents of the class (descriptive power), while the other tends to be higher if the term appears only in documents of the class (discriminating power). Another scheme uses a *relevancy frequency factor* (RF) [15] that takes into account term distribution across classes. According to this scheme, the higher the concentration of high-frequency terms in the positive category than in the negative one, the greater the contribution to classification. Domeniconi et al. (2015) propose a supervised variant of IDF called *inverse document frequency excluding category* (IDFEC). Similar to IDF, IDFEC penalizes frequent terms, but different form IDF it avoids penalizing those terms occurring in several documents belonging to the same class. Another variant also proposed in [7] results from combining IDFEC and RF, resulting in the IDFEC_B scheme.

Table 1 shows the definitions of the main scores presented above using the following notation [7, 14]:

- A denotes the number of documents that belong to class c_k and contain term t_i.
- B denotes the number of documents that belong to class c_k but do not contain the term t_i.
- C denotes the number of documents that do not belong to class c_k but contain the term t_i.
- D denotes the number of documents that do not belong to class c_k class and do not contain the term t_i.
- N denotes the total number of documents in the collection (i.e., $N = A + B + C + D$).

Note that some formulations include the expression $\max(X, 1)$ to prevent the possibility of undefined values, such as divisions by zero or $\log(0)$.

3 A Novel Supervised Term-Weighting Score

Based on the idea that class labels convey useful information for term weighting and on the fact that the importance of a term in a topic or domain depends on the specific objectives at hand (e.g., attaining high recall, high precision or both), we distinguish two relevancy scores. The first score represents the importance of a term to describe the class or topic, and we refer to it as *descriptive relevance* (DESCR). Given a term t_i and a class c_k the DESCR score is expressed as:

$$\text{DESCR}(t_i, c_k) = \frac{|d_j : t_i \in d_j \wedge d_j \in c_k|}{|d_j : d_j \in c_k|},$$

which is equivalent to $A/(A + B)$, using the notation adopted in the previous section. The descriptive relevance of a term in a class stands for a simple idea: those terms that occur in many documents of a given class are good descriptors of that class. As a consequence, we compute it as the portion of documents in the class that contain the given term.

Name	Formulation
TGF	$A + C$
IDF	$\log(N/(A + C))$
TGF*	A
MI	$\log((N \times \max(A, 1))/((A + B)(A + C)))$
χ^2	$N((AD - BC)^2/((A + C)(B + D)(A + B)(C + D)))$
OR	$\log((\max(A, 1) \times D)/\max(B \times C, 1))$
IG	$(A/N)\log(\max(A, 1)/(A+C)) - ((A+B)/N)\log((A+B)/N) + (B/N)\log(B/(B+D))$
GR	$IG/(-((A + B)/N)\log((A + B)/N) - ((C + D)/N)\log((C + D)/N))$
GSS	$\log(2 + ((A + C + D)/(\max(C, 1))))$
Prob	$\log(1 + (A/B)(A/C))$
RF	$\log(2 + (A/\max(C, 1))$
IDFEC	$\log((C + D)/\max(C, 1))$
TGF-IDFEC	$(A + C)(\log((C + D)/\max(C, 1)))$
TGF*-IDFEC	$A \times (\log((C + D)/\max(C, 1)))$
IDFEC_B	$\log(2 + (A + C + D)/(\max(C, 1)))$

Table 1: Definitions of term weighting schemes.

The second relevancy score represents the importance of a term to discriminate a class or topic, and we call it *discriminative relevance*. For a term t_i and a class c_k the DISCR score is expressed as:

$$\text{DISCR}(t_i, c_k) = \frac{|d_j : t_i \in d_j \wedge d_j \in c_k|}{|d_j : t_i \in d_j|},$$

which is equivalent to $A/(A + C)$. The discriminative relevance of a term in a class is based on the idea that a term is a good discriminator of a class if it tends to occur only in documents of that class. We compute it as the portion of documents that contain the given term that belong to the class. The DESCR and DISCR scores can be seen as the supervised versions of the semi-supervised techniques proposed in [21, 22] to compute the descriptive and discriminative power of a term in a topic.

To better understand the notions of descriptors and discriminators, the best terms based on these metrics were analyzed for the Economy domain using an expert manually labeled collection (described in section 4.1). The terms with the highest descriptive power (DESCR > 0.2) and the ones with the highest discriminating power (DISCR > 0.85) are shown in the word clouds of figures 1 and 2, respectively. It is possible to observe that descriptors are terms that are common in the Economy domain, while discriminators are more specific terms that can be found in particular areas in the Economy domain.

We propose to combine the DESCR and DISCR scores by means of the following general term relevancy formula:

$$\text{FDD}_\beta(t_i, c_k) = (1 + \beta^2)\frac{\text{DISCR}(t_i, c_k) \times \text{DESCR}(t_i, c_k)}{(\beta^2 \times \text{DISCR}(t_i, c_k)) + \text{DESCR}(t_i, c_k)}.$$

The FDD_β measure is derived from the F_β formula traditionally used in information retrieval to give β times more importance to recall than to precision:

$$\text{F}_\beta(t_i, c_k) = (1 + \beta^2)\frac{\text{precision}(t_i, c_k) \times \text{recall}(t_i, c_k)}{(\beta^2 \times \text{precision}(t_i, c_k)) + \text{recall}(t_i, c_k)}.$$

By using a β value higher than 1 in the FDD_β function we can weight descriptive relevance higher than discriminative relevance (by placing more emphasis on terms that help achieving good recall) while a β smaller than 1 weights descriptive relevance lower than discriminative relevance (by placing more emphasis on terms that help achieving good precision).

We will show next that FDD_β, can serve the purpose of approximating term relevancy in a topic. This score can be computed for any collection of documents labeled as relevant or irrelevant to the given topic.

We will show next that despite the simplicity of the FDD_β score, it is highly effective both as an estimator of expert assessments of relevance and for guiding the selection of good query terms. In particular, we will show how the tunable parameter β offers a means to favor different objectives in the information retrieval task.

Figure 1: Word cloud based on the top-ranked terms according to the DESCR weighting scheme (only words with DESCR > 0.02 are shown).

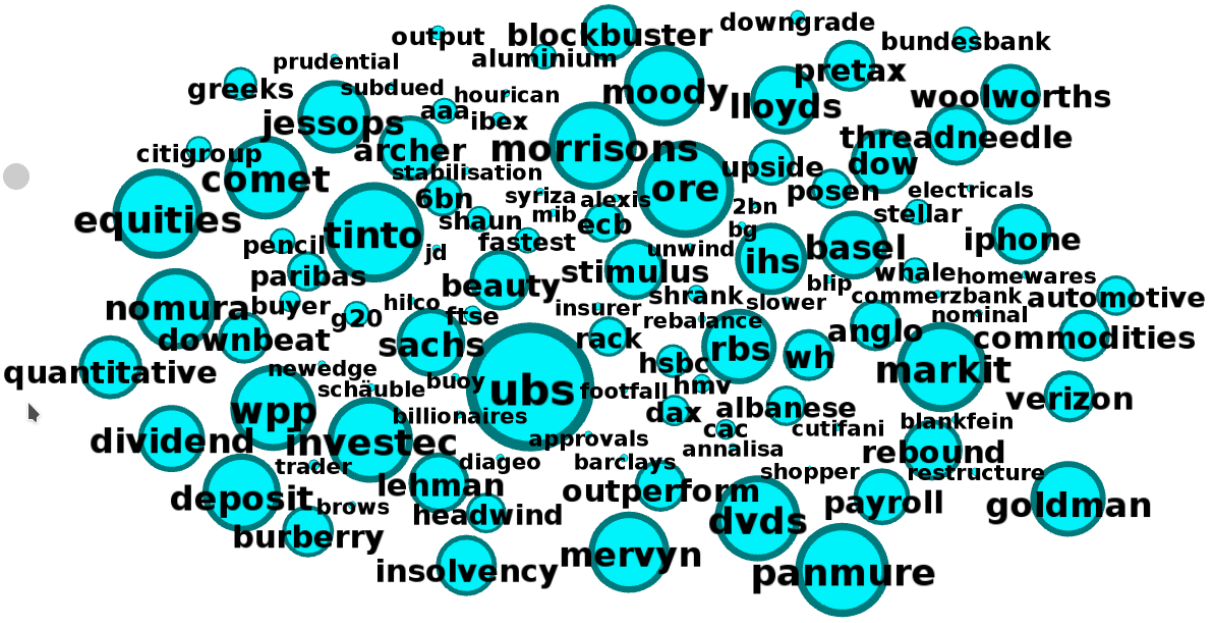


Figure 2: Word cloud based on the top-ranked terms according to the DISCR weighting scheme (only words with DISCR > 0.85 are shown).

4 Evaluation

The goal of this section is to compare the FDD_β weighting score against other term-weighting schemes. The evaluation comprises a human-subject study and an experiment for assessing the effectiveness of the evaluated techniques in information retrieval. The evaluations were carried out using a labeled collection of news articles and human subjects' relevance assessments, as described next. The labeled collection of news articles and the

human subjects' relevance assessments are available for download at http://ir.cs.uns.edu.ar/datasets.

4.1 Data Collection

The *The Guardian* newspaper (https://www.theguardian.com/) was selected as a source to collect a set of digital news. *The Guardian* is a British daily newspaper with an open platform that allows accessing over 1.9 million pieces of content, including full-text news articles. A simple Python script was developed to collect news articles through an API provided by the newspaper. Only news coming from the Politics, World news, Business and Society sections were collected. Although several fields are available for each news article, only the news titles and full body text were used. A simple preprocessing step was carried out to eliminate stopwords and punctuation marks, as well as to transform the text into a sequence of lowercase terms. A total of 1689 news articles corresponding to January 2013 were manually labeled by two experts in Economy as relevant (537) or irrelevant (1152) to the economy. News were considered relevant to the economy if the described event has some impact on the economy.

To complete the labeling task, both experts read the news articles and agreed on whether each of them was relevant or not. It is worth mentioning that the manual labeling stage was important due to the fact that news identified by the experts as relevant to the economy do not exactly correspond to those from the "Business" section (418 out of 512) but also some of them were in the Politics (39 out of 290), World news (43 out of 650), and Society (37 out of 237) sections.

To better understand the discrepancies between the news in the "Business" section and what was identified by the experts as relevant to the economy we present in table 2 some examples of news titles that were labeled by the experts. In these example it is possible to see that many news that belong to the "Business" section have to do with the *World Economic Forum* that took place in Davos in 2013. While this social event may be of interest to the Business community it was considered by the experts as non-relevant to the economy. Similarly, other news that describe social events related to the area of Economy were labeled as non-relevant by the experts. On the other hand, the analysis of news outside the "Business" section allowed to identify some news with impact on local and/or global economies. This is the case of many news in the "Politics" section that involved the announcement of political decisions of countries. Other news in the "Society" section that were identified as relevant to the economy include news about retirements, social benefits, public health, among other topics. While these news have a focus on society, and hence belong to the "Society" section, many of them have a direct impact on the economy. As is illustrated by the selected examples, the "World news" section also included several news from countries different from Great Britain that were identified by the expert as news with an impact on the economy.

The collection of 1689 expert-labeled news articles was used as the training set. Also, a reduced set consisting of 100 expert-labeled news articles (not included in the training set) was used as the validation set. The total number of terms in these news articles is 38511. However, to reduce the dimensionality of the dataset only those terms that occur in at least six news articles were considered, resulting in a set of 10373 terms.

4.2 Validation by User Study

Eight volunteer subjects were recruited for an experiment conducted online. The group of subjects included four volunteers with no background in Economy and four others with a Ph.D. degree in Economy. We refer to the first group as *non-experts* and to the second group as *experts*. The motivation for examining and comparing the assessments of both expert and non-expert users was to determine if both populations exhibit different behaviors and to evaluate the discrepancy between our tool and the two groups. A set of 50 terms (10 lists of 5 terms each) and another set of 100 terms (20 lists of 5 terms each) were strategically selected from the 10373 terms of the dataset. The selection was made based on the distribution of term frequency in light of Zipf's law. The goal was to avoid providing low-frequency terms (which are many) more chances of being selected than high-frequency terms (which are a few). To complete an initial parameter-adjusting stage, two of the experts were asked to agree on the economic relevance of each of the words from the 50-term set. The experts were asked to rate these terms with a score ranging from 0 (economic irrelevant) to 5 (economic very relevant). We used these ratings and the labeled collection to learn the best β value for the FDD_β method. As can be seen in figure 3 the highest Pearson correlation between the expert ratings and the FDD_β values was 0.797671, which was achieved for $\beta = 0.477$.

To complete the validation stage we asked the eight volunteer subjects to rate the 100 terms using a 0-5 scale, and we computed DESCR, DISCR, $\text{FDD}_{0.477}$ and the 15 weighting schemes listed in table 1 for these terms. In the first place, we tested the level of agreement between pairs of users belonging to the non-expert group and between pairs of users in the expert group. Table 3 presents the means and standard deviations obtained as a result of such analysis. It is possible to observe that there is a high level of agreement in both groups, being this agreement higher in the expert group.

Table 2:

News in the Business section but not relevant to the economy		
Title	Date	Section
Jessops goes into administration: staff and customers react	2013-01-10	Business
How to spot a fake indie business	2013-01-03	Business
Women's rights activists protest at Davos - in pictures	2013-01-26	Business
Davos diary: Paul Coelho becomes most retweeted attendee	2013-01-24	Business
Davos 2013: day two - as it happened	2013-01-24	Business
News outside the Business section but relevant to the economy		
Title	Date	Section
Spain's economy shrinks again and remains deep in recession	2013-01-30	World news
Britain leaving EU major threat to global economy, says Sir Martin Sorrell	2013-01-23	Politics
Benefits and child credits squeeze pushes 200,000 children into poverty	2013-01-17	Society
Cameron faces unfriendly fire from military chiefs over defence budget	2013-01-31	Politics
News in the Business section and relevant to the economy		
Title	Date	Section
Shell dividend pleases shareholders but profits disappoint City	2013-01-31	Business
US recovery stalls after first quarter of negative growth in three years	2013-01-30	Business
US economy shrinks unexpectedly despite improving job market	2013-01-30	Business
UK GDP shrank by 0.3% in fourth quarter	2013-01-25	Business
News outside the Business section and not relevant to the economy		
Title	Date	Section
Four US states considering laws that challenge teaching of evolution	2013-01-31	World news
Syria may respond to Israeli air strike, says ambassador	2013-01-31	World news
David Cameron arrives in Libya on surprise visit	2013-01-31	Politics
Queen Beatrix of the Netherlands - in pictures	2013-01-28	World news
Hugo Chávez fights for life as supporters pray in Venezuela	2013-01-04	World news
Man arrested in East Sussex over Nepal war crimes	2013-01-03	World news
Criminals should spend longer in jail, says Chris Grayling	2013-01-05	Society

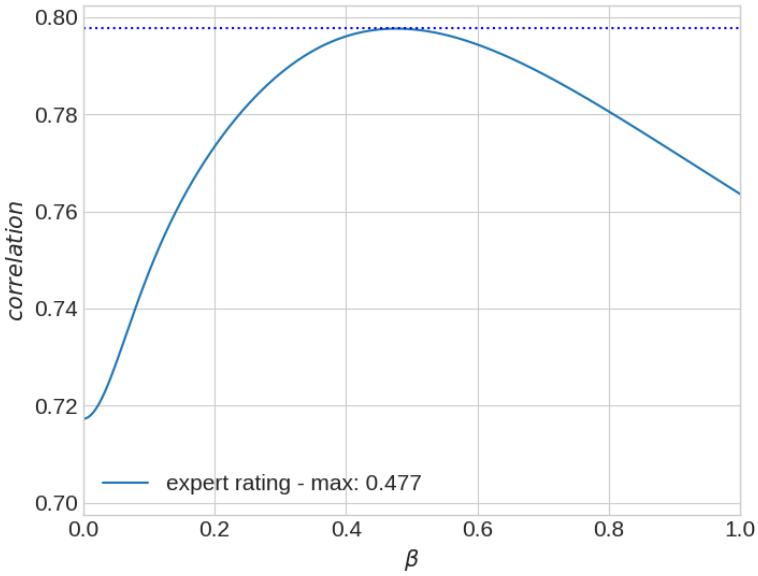


Figure 3: Learning the optimal β value (maximum correlation equal to 0.797671 for $\beta = 0.477$).

non-expert	experts
$\mu = 0.839475$, $\sigma = 0.037791$	$\mu = 0.876390$, $\sigma = 0.009438$

Table 3: Means (μ) and standard deviations (σ) of correlations to test agreement among non-experts and among experts.

Table 4 presents results on the level of agreement between the two different groups of users (non-experts and experts), and on the level of agreement between each of these groups (non-experts and experts) and $\text{FDD}_{0.477}$. The level of agreement is given by the averaged Pearson correlation coefficients.

non-experts and experts	non-experts and $\text{FDD}_{0.477}$	experts and $\text{FDD}_{0.477}$
$\mu = 0.80383$, $\sigma = 0.053205$	$\mu = 0.685598$, $\sigma = 0.054969$	$\mu = 0.752352$, $\sigma = 0.018904$

Table 4: Means (μ) and standard deviations (σ) of correlations computed between non-experts and experts, non-experts and $\text{FDD}_{0.477}$, and experts and $\text{FDD}_{0.477}$.

Finally, to compare the effectiveness of the weighting schemes as predictors of subjects' judgments of term relevancy we computed the Pearson correlation coefficients between the averaged ratings assigned by the subjects and those computed by each of the weighting schemes. Table 5 summarizes these correlations. The reported values correspond to the correlations between each of the methods and the different groups of users. In all these cases we observe that $\text{FDD}_{0.477}$ outperforms the other methods, being TGF*-IDFEC the second most effective one in estimating human subjects' relevance assessments.

4.3 Retrieval Effectiveness

In this section, we analyze the performance of FDD_β as a mechanism for query-term selection, and we compare it with other state-of-the-art weighting schemes. In the first place, the training set described in section 4.1 was used to select the top-rated terms based on FDD_β by assigning different values to the parameter β. Simple queries were generated using the selected terms and then evaluated by means of the classical recall, precision and F_1 metrics. The results are shown in figure 4. As expected, the highest recall using the FDD_β-based term selection mechanisms is obtained with larger values of β while the highest precision is obtained for smaller values. Note, for instance, that terms such as `uk` occur often in relevant news articles given the fact that news were collected from a British newspaper. As a result, the term `uk` results in a high-recall query. However, `uk` is not a good discriminator for the Economy domain, resulting in a low-precision query. On the other hand, terms such as `adp`,

Method	non-expert (averaged)	expert (averaged)	non-expert and expert (averaged)
TGF	0.283553	0.365037	0.332324
IDF	-0.488816	-0.563704	-0.539138
TGF*	0.574110	0.642607	0.623198
MI	0.697053	0.659659	0.694604
χ^2	-0.164537	-0.087771	-0.128992
OR	0.432627	0.306599	0.378188
IG	0.663296	0.705736	0.701123
GR	0.663296	0.705736	0.701123
GSS	0.722761	0.757015	0.757807
Prob	0.654187	0.697007	0.691990
RF	0.472824	0.407394	0.450543
IDFEC	-0.226397	-0.325872	-0.283050
TGF-IDFEC	0.603975	0.676551	0.655882
TGF*-IDFEC	0.721871	0.774026	0.766110
IDFEC_B	-0.221061	-0.320304	-0.277466
DESCR	0.574110	0.642607	0.623198
DISCR	0.662481	0.610804	0.651848
$FDD_{0.477}$	**0.735456**	**0.791969**	**0.782264**

Table 5: Correlations between methods and ratings obtained by averaging non-expert, expert and all human subjects' scores.

`jp`, `ubs`, `forecasts` and `ftse` are not good descriptors but tend to occur only in relevant news articles. This means that they are good discriminators, offering a mechanism to ensure high precision, although usually at the expense of low recall. Other terms, such as `sales`, `growth` and `business` achieve a balance between descriptive and discriminative relevance, resulting in a good F_1 score.

The query term with the highest F_1 score is `growth`, which is the term achieving the best FDD_β for a range of β values that begins approximately at 0.4 and ends close to 1.2. Note that this range includes 0.477, which is the value that yields the highest correlation between FDD_β scores and experts' relevance assessments. Based on this preliminary analysis the best FDD_β achieves an F_1 score as high as the one obtained with the two most effective state-of-the-art weighting schemes (TGF-IDFEC and TGF*-IDFEC). The top-rated term according to the three weighting schemes is `growth`. It is interesting to note that for small β values FDD_β outperforms these two methods in terms of precision while for large β values FDD_β outperforms these two methods in terms of recall.

The validation set described in section 4.1 was used to determine if the best queries identified using the training set were effective on a different set. The resulting recall, precision and F_1 metrics computed on the validation set are shown in figure 5. Given that the validation set was small, some of the most discriminating terms identified during the training stage (`adp` and `ubs`) were absent from the validation set, resulting in an empty answer set when used as query terms. However, those terms with a good balance between descriptive and discriminative relevance (`sales`, `growth` and `business`) achieve the highest F_1 scores when used as query terms on the validation set. This preliminary analysis indicates that the proposed method does not overfit the training data.

To further investigate into the effectiveness of FDD_β, we used the top-ranked terms based on different β values to formulate different types of queries. The evaluated queries for FDD_β include single-term queries (FDD_β), disjunctive queries with two terms (FDD_β (OR(2))), disjunctive queries with three terms (FDD_β (OR(3))), conjunctive queries with two terms (FDD_β (AND(2))) and conjunctive queries with three terms (FDD_β (AND(3))). For comparison purposes, we used TGF*-IDFEC, which based on our previous analysis represents one of the most effective state-of-the-art weighting schemes, and proceeded to formulate different queries based on the top-ranked terms according to this scheme. In this sense, the evaluated queries for TGF*-IDFEC include single-term queries (TGF*-IDFEC), disjunctive queries with two terms (TGF*-IDFEC (OR(2))), disjunctive queries with three terms (TGF*-IDFEC (OR(3))), conjunctive queries with two terms (TGF*-IDFEC (AND(2))) and conjunctive queries with three terms (TGF*-IDFEC (AND(3))).

Figure 6 shows the effectiveness of these queries on the training set based on recall, precision and F_1. These

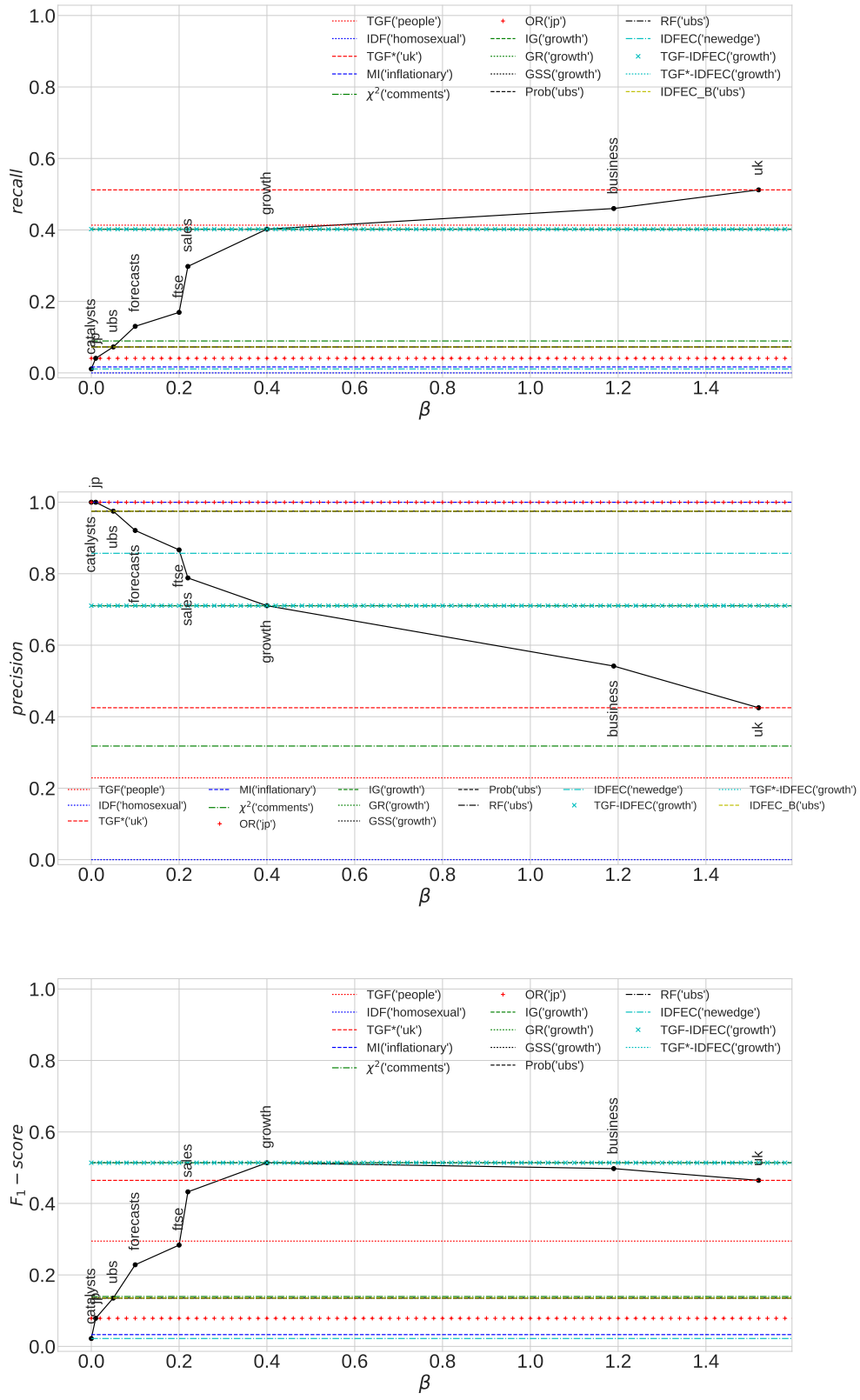


Figure 4: Effectiveness on the training set of queries generated based on term weighting schemes. The black solid curve corresponds to the effectiveness of query terms selected using FDD_β on the training with different β values.

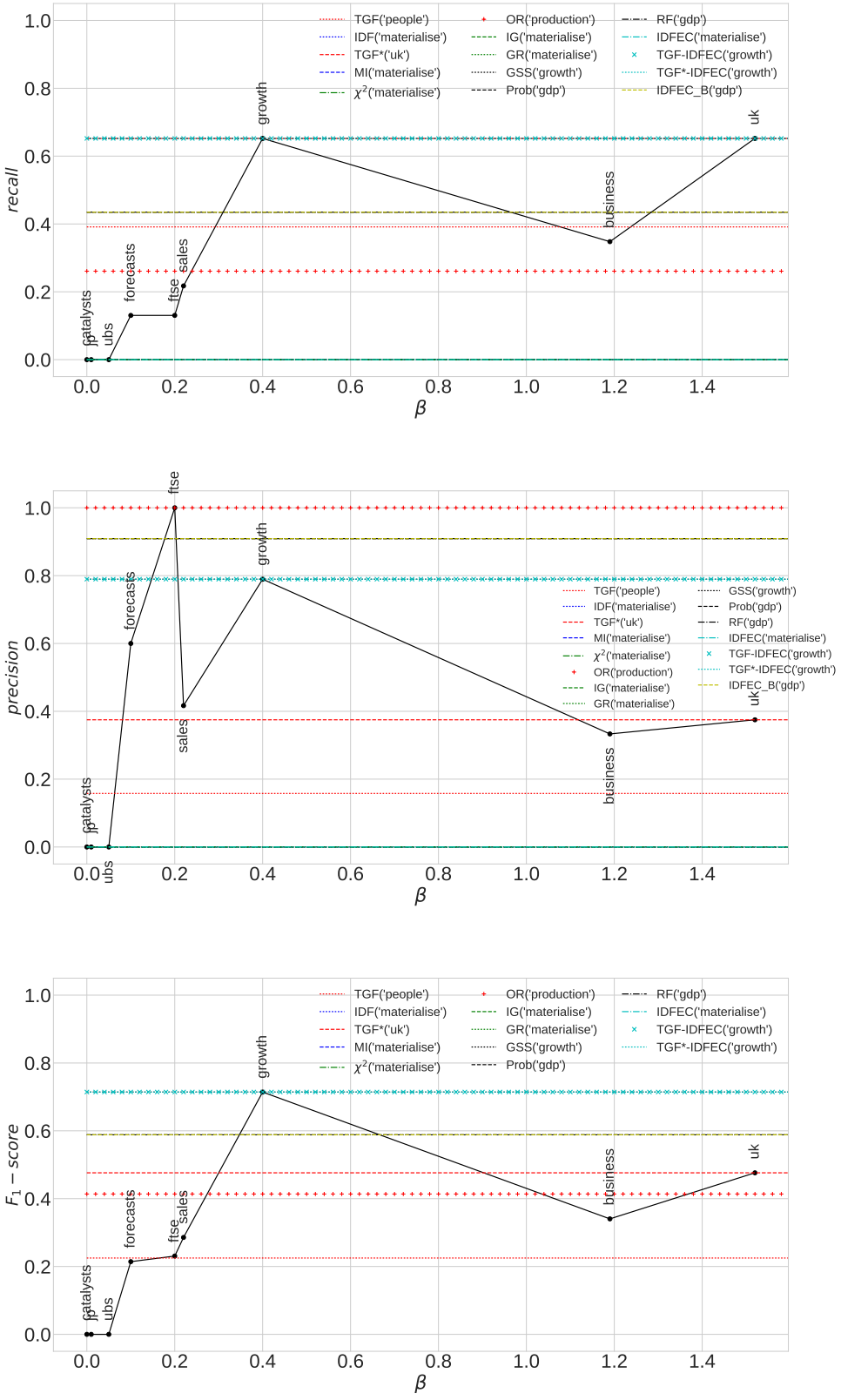


Figure 5: Effectiveness on the validation set of queries generated based on term weighting schemes. The black solid curve corresponds to the effectiveness of query terms selected using FDD_β on the training set with different β values

results indicate that disjunctive queries with three terms selected based on FDD_β (FDD_β (OR(3))) achieve the same F_1 as the best TGF*-IDFEC-based query when the FDD_β-based queries take a β-value in an interval that includes 0.477 (which is the β with the highest correlation between FDD_β scores and experts' relevance assessments). Also, it is interesting to note that for high β values, the "FDD_β (OR(3))" queries outperform all the TGF*-IDFEC queries in terms of recall, while for small β values several combinations of FDD_β-based queries outperform the TGF*-IDFEC-based queries in terms of precision.

Figure 7 shows the performance of the queries described above on the validation set. These results show that for some ranges of β, the single-term queries (FDD_β) and some of the disjunctive queries with three terms (FDD_β (OR(3))) achieve an F_1 score equal to that achieved by the best TGF*-IDFEC-based query. Once again, some of the queries generated based on the "FDD_β (OR(3))" scheme outperform all other queries in terms of recall. On the other hand several FDD_β and TGF*-IDFEC-based schemes achieve the highest possible precision value (1.0). Once again, the results indicate that the proposed query generation schemes do not overfit the training data.

5 Application in Variable Extraction, Modeling and Retrieval

As can be seen in the reported results, FDD_β performs consistently well, not only as an estimator of human subjects' relevance assessments but also as a method for guiding the selection of good query terms. This opens numerous opportunities for applying the proposed scheme on different scenarios. This section describes some possible applications that may benefit from the proposed term-weighting technique.

5.1 Building Prediction Models

The proposed technique can be used to extract variables from digital media with the ultimate goal of building models of prediction, explanation and description. Figure 8 shows a word-cloud visualization with the top-ranked terms based on the training data using $FDD_{0.477}$ as weighting scheme. To avoid overcharging the figure only terms with $FDD_{0.477} > 0.6$ are shown.

It is worth mentioning that the proposed FDD_β measure can be computed not only for words in a collection (as illustrated so far) but also on other types of lexical units, such as stems, n-grams, named entities, noun phrases, among others. Figure 9 presents a word cloud with the top-ranked named entities based on $FDD_{0.477}$ identified in the manually labeled data collection used in our previous analysis. The *Stanford Named Entity Recognition* tagger (Stanford NER) [10] was used to identify these named entities.

A subsequent modeling step would be to identify different types of dependency relations between these variables. For instance, some *causal relations* that can be recognized are `investment-growth-gdp`, `spending-market-recovery` and `sales-companies-investment-gdp`. Other types of relations, such as *close associations* are illustrated by `credit-debt-banks`, `recession-decline`, `trading-stock-ftse` and `debt-bank-investors-trading-recession`. A possible *simultaneity relation* is given by *market-prices*. It is also interesting to note that `christmas`, one of the words selected by the method, may capture *seasonality in a casual series*. Automatically identifying these types of relations is a challenging problem that we plan to address as part of our ongoing research work. In particular, we plan to investigate into the problem of finding causal relations with the purpose of automatically building different types of networks, such as Bayesian networks [26].

5.2 Supporting Knowledge Modeling

Building knowledge models is a difficult and costly task. There are several initiatives aimed at providing intelligent support to facilitate the construction of knowledge models, as is the case of the family of intelligent suggesters for concept mapping described in [17, 20].

Concept mapping [24] is a vehicle for knowledge modeling that was first proposed in education, to enable students to externalize their knowledge by constructing a graphical representation of concepts and their relationships. Concept mapping systems have been used by users ranging from elementary school students to scientists to support generation, storage of, and access to concept maps in electronic form. In addition to providing basic operations needed to draw concept maps, these systems can be augmented with methods aimed at facilitating knowledge extension. In particular, the automatic identification of terms relevant to the modeled domain allows to extend a knowledge model beyond information that has already been captured. The visualizations presented in figures 8 and 9 can support the construction of concept maps, by helping in the process of choosing relevant terms in a particular domain, which is typically the initial step in any knowledge modeling task.

Concept maps are usually organized in a hierarchical fashion, where more general terms tend to appear on the top of the concept map (i.e., in or close to the root node), while more specific terms tend to occur toward the bottom (i.e, in or close to the leaf concepts). The term-weighting method proposed in this work may offer a novel solution to the problem of identifying different terms and entities that can be suggested for addition at

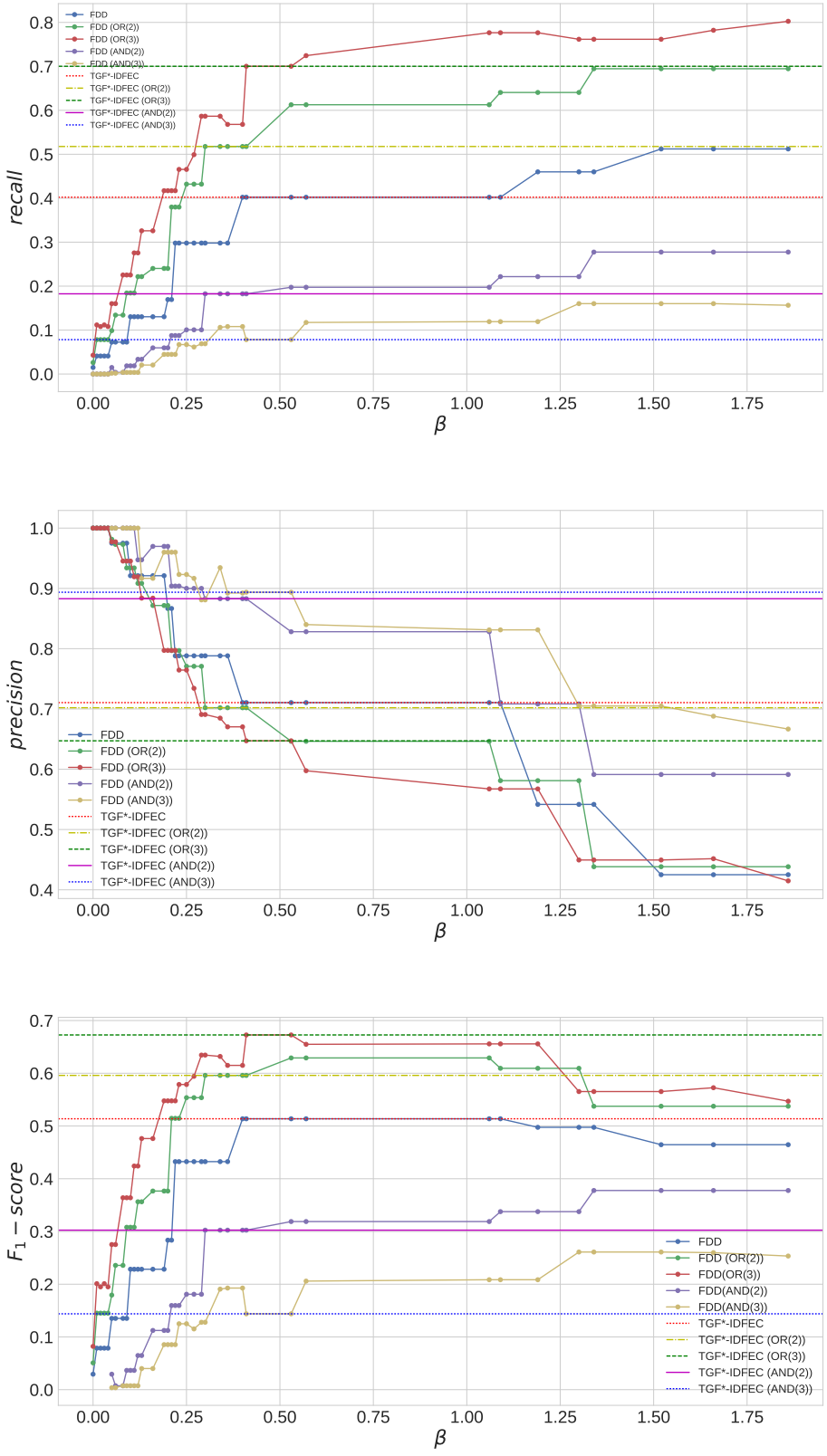


Figure 6: Effectiveness on the training set of one-, two- and three-term conjunctive and disjunctive queries generated based on the FDD_β and TGF*-IDFEC term weighting schemes. The solid curves correspond to the effectiveness of query terms selected using FDD_β with different β values.

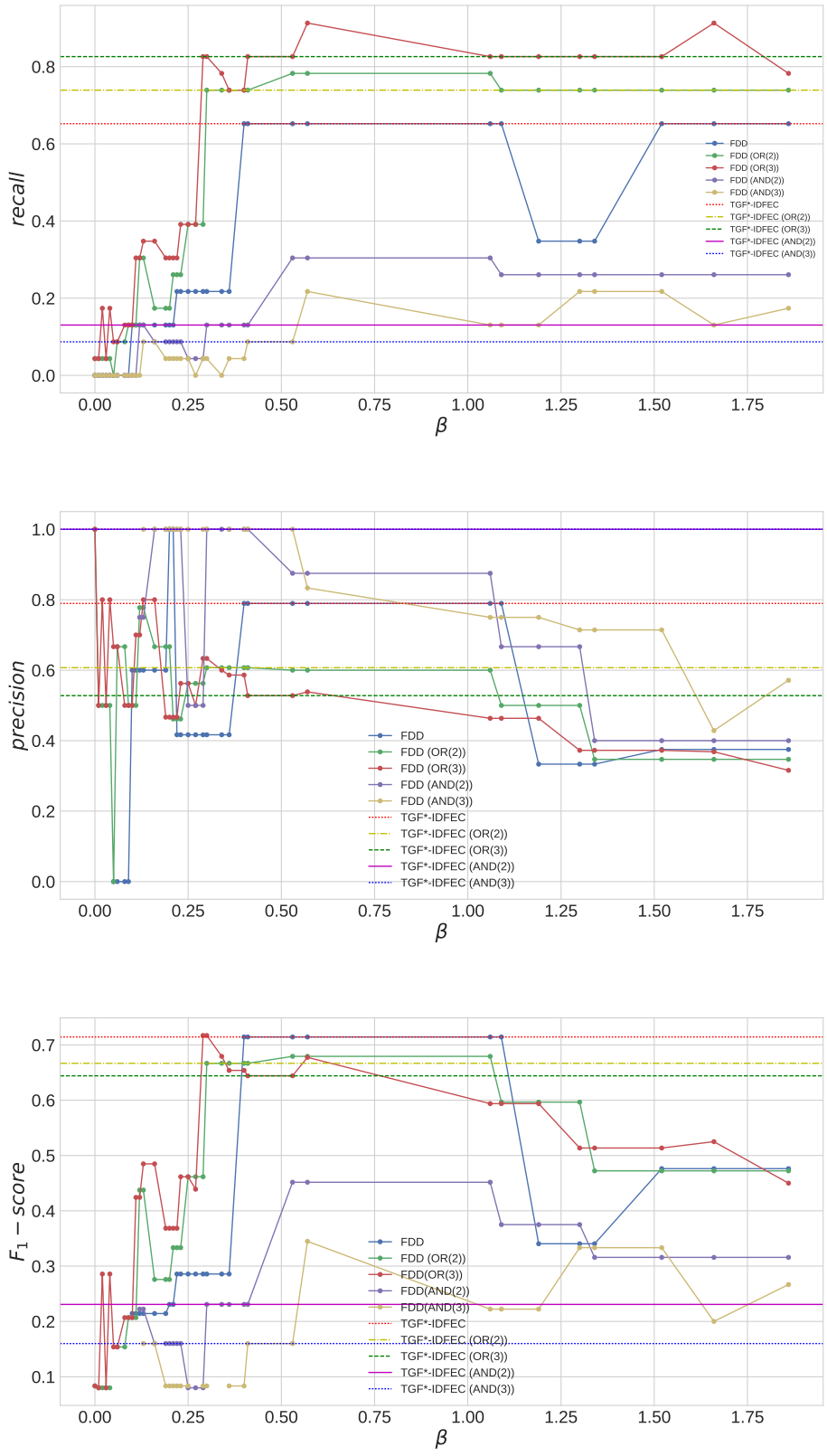


Figure 7: Effectiveness on the validation set of one-, two- and three-term conjunctive and disjunctive queries generated based on the FDD_β and TGF*-IDFEC term weighting schemes. The solid curves correspond to the effectiveness of query terms selected using FDD_β with different β values.

Figure 8: Word cloud based on the top-ranked terms according to the $FDD_{0.477}$ weighting scheme (only words with $FDD_{0.477} > 0.6$ are shown).

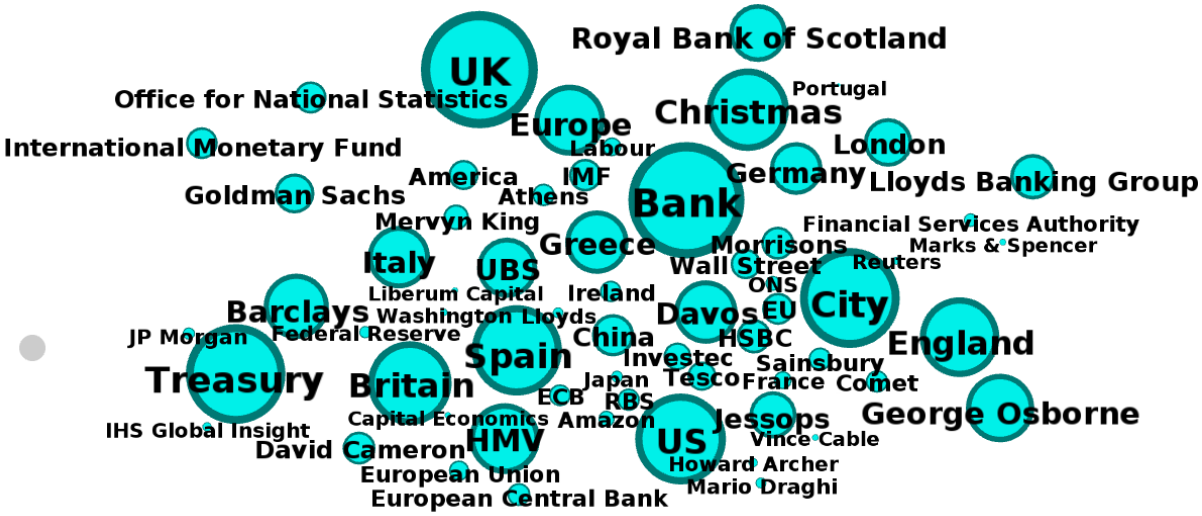


Figure 9: Word cloud based on the top-ranked named entities according to the $FDD_{0.477}$ weighting scheme (only words with $FDD_{0.477} > 0.15$ are shown).

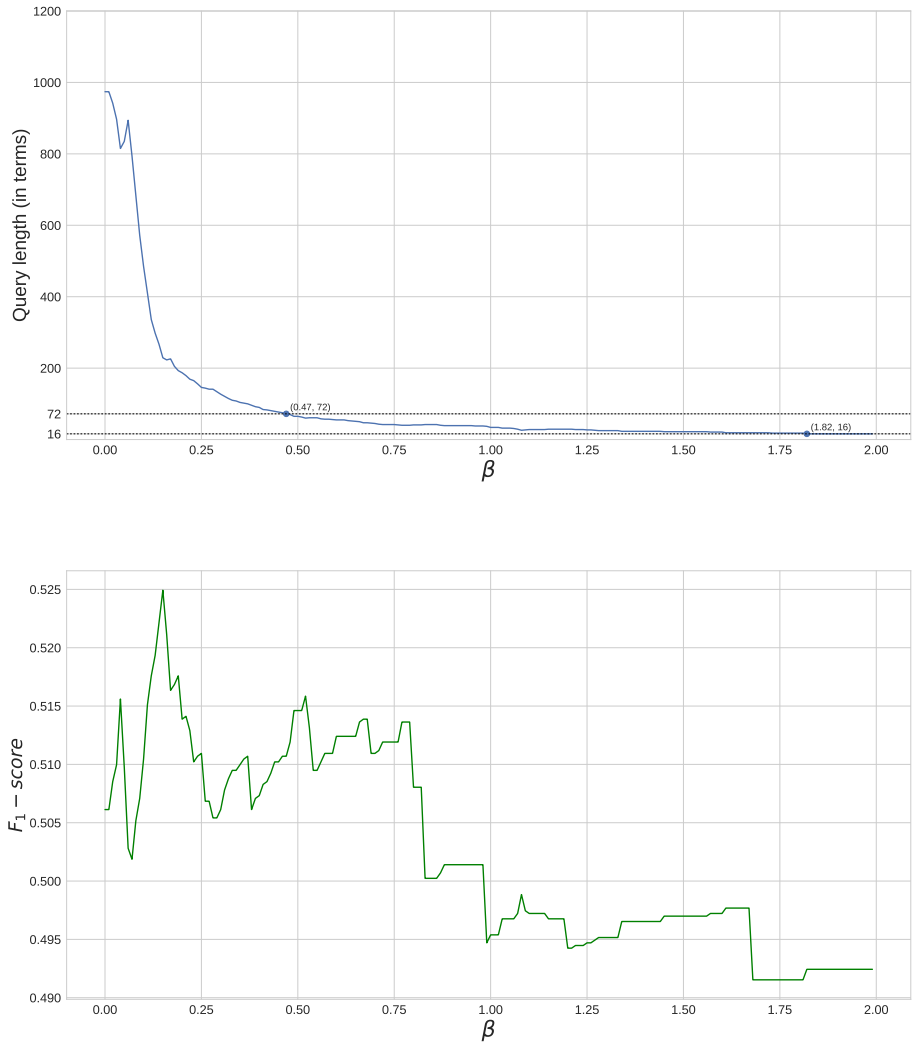


Figure 10: Number of top-ranked terms based on different β values needed in a disjunctive query to achieve total recall (i.e., recall =1) on the training set (top) and F_1 values achieved for these queries on the same set (bottom).

different levels of a concept maps. Since descriptors tend to represent terms describing a general domain or topic while discriminators tend to be specific terms, the use of an adjustable β parameter in the FDD_β measure can help focus the term selection process to favor generality or specificity, depending on the user needs.

5.3 Achieving Total Recall

The problem of total-recall (or high-recall) retrieval [1, 12, 27] is to find all (or nearly all) relevant documents for a search topic. As opposed to those scenarios where the goal is to retrieve a few high quality relevant documents (e.g., answering a specific consultation need with high precision), total recall scenarios are those where the focus is on retrieving all relevant documents, without a significant loss in precision (e.g., collecting all documents relevant to a topic or domain to build a topical or domain-based web portal).

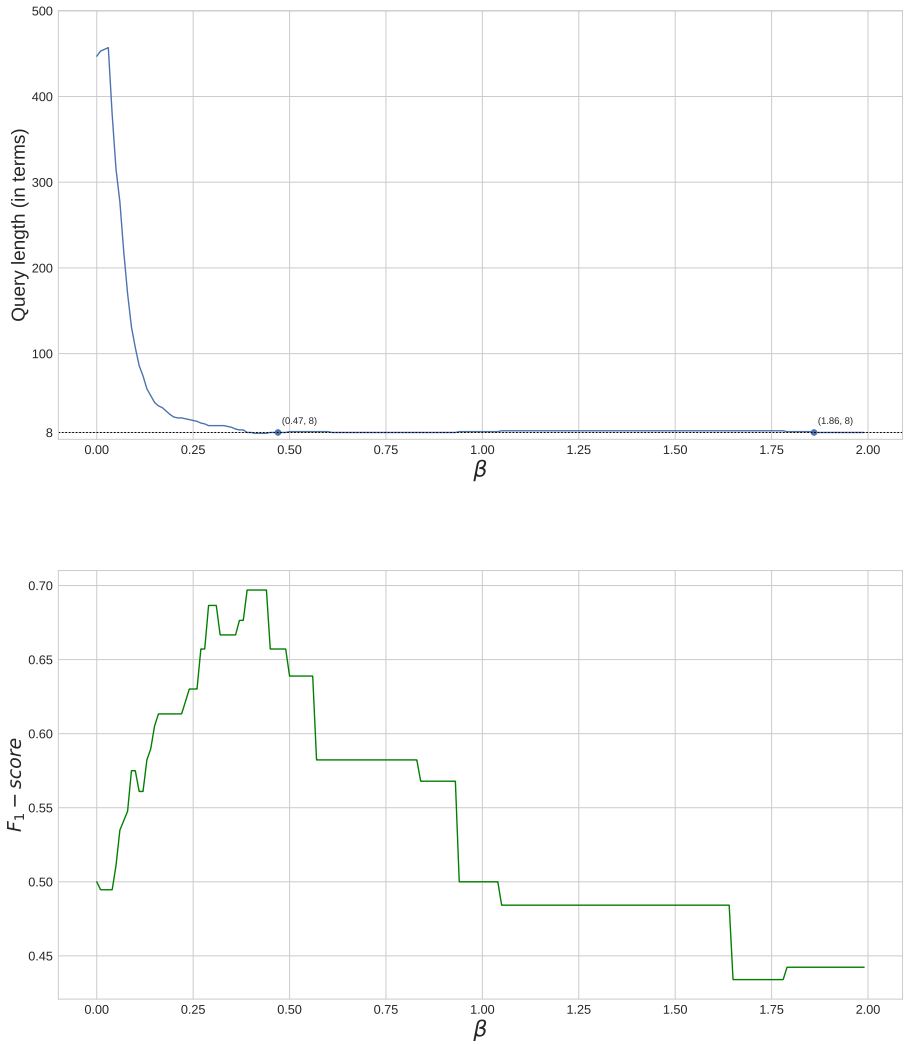


Figure 11: Number of top-ranked terms (learned from the training set) based on different β values needed in a disjunctive query to achieve total recall (i.e., recall =1) on the validation set (top) and F_1 values achieved for these queries on the same set (bottom).

A question that arises when addressing the total recall problem is how many terms are necessary to build a disjunctive query that achieves 100% recall. To look into this question we used our training data collection to incrementally construct disjunctive queries by adding the top-ranked terms based on their FDD_β score for different β values. As discussed earlier, terms with high descriptive power are those terms that tend to occur often in relevant documents. Hence, we expected that when incrementally longer queries are built with a focus on the descriptive power of the query terms (i.e., large β values) it should be possible to construct shorter high-recall queries than when the focus is placed on the discriminative power of the terms (i.e., small β values). This intuition is verified by the results shown in figure 10, where it is possible to see that as β increases, the number of terms needed to achieve total recall reduces significantly.

A similar analysis on the validation set is presented in figure 11. Once again the results indicate that the larger the β value in the selection of top-ranked terms the smaller is the number of terms needed to form a total-recall query. Note that the query length required to achieve total recall is highly dependent on the number of relevant documents in the collection. Hence, total-recall queries for the testing set are significantly shorter than total-recall queries for the training set.

To analyze how total-recall queries impact the F_1 scores we also report F_1 for the evaluated queries both on the training set (bottom of figure 10) and the validation set (bottom of figure 11). For both collections it is possible to see that for certain values of β, total recall is possible without a significant loss in F_1. These results point to the potential of the proposed technique as a mechanism to further investigate the total recall problem.

6 Conclusions and Future Work

In this paper we presented a methodology for identifying domain-specific terms. As part of the proposed methodology we defined a novel supervised term-weighting scheme called FDD_β, which is based on the notions of descriptive and discriminative relevance. Preliminary evaluations show that FDD_β achieves good performance as an estimator of human subjects' relevance judgments and as a mechanism for selecting good query terms. Also, it offers the flexibility of adapting to different goals, such as achieving high recall, high precision, or a balance between both. This flexibility represents an important advantage over the analyzed state-of-the-art weighting schemes. In particular, it offers a novel mechanism for query refinement by guiding the selection of more descriptive query terms to retrieve more general results when initial results are to narrow. Similarly, it can help identifying more discriminative query terms to retrieve more specific results when initial results are too broad.

While the analysis presented here is focused on retrieval, the proposed term-weighting technique may also be applied to classification. This will open new challenges such as analyzing strategies for combining the FDD_β score with local weighting schemes, such as TF.

The proposed technique was evaluated on the economic domain with promising results and we anticipate that it will also achieve good performance on other domains. Also, we plan to test FDD_β on specific topics (as is the case of the topic of a news article), as opposed to general domains (as is the case of Economy). Another important future task will be to validate FDD_β on larger data sets, such as those available as part of the TREC collection (https://trec.nist.gov/data/test_coll.html). Finally, we plan to investigate the definition of non-supervised versions of the proposed weighting techniques, where topic relevance will be approximated by clustering and other non-supervised approaches.

Acknowledgment

This work was supported by CONICET (PIP 11220120100487), MinCyT (PICT 2014-0624) and Universidad Nacional del Sur (PGI-UNS 24/N029).

References

[1] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D Smucker, Gordon V Cormack, and Maura R Grossman. A system for efficient high-recall retrieval. In *SIGIR*, pages 1317–1320, 2018.

[2] Kewen Chen, Zuping Zhang, Jun Long, and Hao Zhang. Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications*, 66:245–260, 2016.

[3] Franca Debole and Fabrizio Sebastiani. Supervised term weighting for automated text categorization. In *Text mining and its applications*, pages 81–97. Springer, 2004.

[4] C Deisy, M Gowri, S Baskar, SMA Kalaiarasi, and N Ramraj. A novel term weighting scheme midf for text categorization. *Journal of Engineering Science and Technology*, 5(1):94–107, 2010.

[5] Zhi-Hong Deng, Shi-Wei Tang, Dong-Qing Yang, Ming Zhang, Xiao-Bin Wu, and Meng Yang. A linear text classification algorithm based on category relevance factors. In *International Conference on Asian Digital Libraries*, pages 88–98. Springer, 2002.

[6] Zhi-Hong Deng, Kun-Hu Luo, and Hong-Liang Yu. A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, 41(7):3506–3513, 2014.

[7] Giacomo Domeniconi, Gianluca Moro, Roberto Pasolini, and Claudio Sartori. A study on term weighting for text categorization: A novel supervised variant of tf. idf. In *DATA*, pages 26–37, 2015.

[8] MA Fattah and MG Sohrab. Combined term weighting scheme using ffnn, ga, mr, sum, and average for text classification. *International Journal of Scientific and Engineering Research*, 7(8):2031–2040, 2016.

[9] Guozhong Feng, Shaoting Li, Tieli Sun, and Bangzuo Zhang. A probabilistic model derived term weighting scheme for text classification. *Pattern Recognition Letters*, 110:23–29, 2018.

[10] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885. URL `https://doi.org/10.3115/1219840.1219885`.

[11] Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi. Experiments on the use of feature selection and negative evidence in automated text categorization. In *International Conference on Theory and Practice of Digital Libraries*, pages 59–68. Springer, 2000.

[12] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. Trec 2016 total recall track overview. In *TREC*, 2016.

[13] Samer Hassan, Rada Mihalcea, and Carmen Banea. Random walk term weighting for improved text classification. *International Journal of Semantic Computing*, 1(04):421–439, 2007.

[14] Man Lan, Sam-Yuan Sung, Hwee-Boon Low, and Chew-Lim Tan. A comparative study on term weighting schemes for text categorization. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*, volume 1, pages 546–551. IEEE, 2005.

[15] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):721–735, 2009.

[16] Christine Largeron, Christophe Moulin, and Mathias Géry. Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 924–928. ACM, 2011.

[17] David Leake, Ana Maguitman, and Thomas Reichherzer. Experience-based support for human-centered knowledge modeling. *Knowledge-Based Systems*, 68(0):77 – 87, 2014. ISSN 0950-7051. URL `http://dx.doi.org/10.1016/j.knosys.2014.01.013`.

[18] Edda Leopold and Jörg Kindermann. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3):423–444, 2002.

[19] Ying Liu, Han Tong Loh, and Aixin Sun. Imbalanced text classification: A term weighting approach. *Expert systems with Applications*, 36(1):690–701, 2009.

[20] Carlos Lorenzetti, Ana Maguitman, David Leake, Filippo Menczer, and Thomas Reichherzer. Mining for topics to suggest knowledge model extensions. *ACM Transactions on Knowledge Discovery from Data*, 11 (2):23:1–23:30, 2016. ISSN 1556-4681. doi: 10.1145/2997657. URL `http://dl.acm.org/authorize?N37228`.

[21] Carlos M Lorenzetti and Ana G Maguitman. A semi-supervised incremental algorithm to automatically formulate topical queries. *Information Sciences*, 179(12):1881–1892, 2009.

[22] Ana Maguitman, David Leake, Thomas Reichherzer, and Filippo Menczer. Dynamic extraction topic descriptors and discriminators: towards automatic context-based topic search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 463–472. ACM, 2004.

[23] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[24] Joseph D Novak. Concept mapping: A useful tool for science education. *Journal of research in science teaching*, 27(10):937–949, 1990.

[25] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[26] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[27] Adam Roegiest, Gordon V Cormack, Maura R Grossman, and Charles Clarke. Trec 2015 total recall track overview. *Proc. TREC-2015*, 2015.

[28] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[29] Takenobu Tokunaga and Iwayama Makoto. Text categorization based on weighted inverse document frequency. In *Special Interest Groups and Information Process Society of Japan (SIG-IPSJ*. Citeseer, 1994.

[30] CJ Van Rijsbergen, David J Harper, and Martin F Porter. The selection of good search terms. *Information Processing & Management*, 17(2):77–91, 1981.

[31] Suzan Verberne, Maya Sappelli, Djoerd Hiemstra, and Wessel Kraaij. Evaluation and analysis of term scoring methods for term extraction. *Information Retrieval Journal*, 19(5):510–545, 2016.

[32] D. Wang and H. Zhang. Inverse-category-frequency based supervised term weighting schemes for text categorization. *Journal of Information Science and Engineering*, 29(2):209–225, 2013. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-84876262286&partnerID=40&md5=00897fb85a83bfd704cb2428254e1950`.