News Recommendation with CF-IDF+

Emma de Koning, Frederik Hogenboom, and Flavius Frasincar

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

370761ek@student.eur.nl, {fhogenboom, frasincar}@ese.eur.nl

Abstract. Traditionally, content-based recommendation is performed using term occurrences, which are leveraged in the TF-IDF method. This method is the defacto standard in text mining and information retrieval. Valuable additional information from domain ontologies, however, is not employed by default. The TF-IDF-based CF-IDF method successfully utilizes the semantics of a domain ontology for news recommendation by detecting ontological concepts instead of terms. However, like other semantics-based methods, CF-IDF fails to consider the different concept relationship types. In this paper, we extend CF-IDF to additionally take into account concept relationship types. Evaluation is performed using Ceryx, an extension to the Hermes news personalization framework. Using a custom news data set, our CF-IDF+ news recommender outperforms the CF-IDF and TF-IDF recommenders in terms of F_1 and Kappa.

Keywords: News Recommender Sytems, CF-IDF, CF-IDF+

1 Introduction

With the continuously growing amount of information on the Internet, distinguishing relevant from irrelevant matters becomes increasingly difficult. What is supposed to be a trivial task for humans, is exceptionally hard for machines because of the difficulties that need to be faced while automatically comprehending the information content. A common type of information that is searched for on the Web is news, i.e., information describing recent events that is or is not yet (partially) known to a user. Articles on news sites are usually already categorized, but this classification is often too course, and does not match the fine-grained user interests. Instead of treating all visitors equally, an automatic recommendation system could aid individual users in selecting interesting information. Determining the user's preferences would be helpful when optimizing browsing experience on news Web sites, for instance by presenting interesting articles on top, or by filtering RSS feeds on interesting contents.

Recommender systems determining the importance of news are not merely useful for enhancing user experiences, but are of utmost importance in decision support scenarios where fast and accurate news provision is needed for reliable

decisions, such as in (algorithmic) trading. In addition, such systems are important to news providing companies with revenue models based on advertisements. Their earnings are usually based on page clicks, which can be maximized by presenting the most relevant items to a user. The visitor is more likely to stay on such sites, yielding more clicks and thus more advertisement earnings.

A vast amount of research has been conducted related to algorithms and systems for news recommendation [4–6, 10, 13, 14]. In general, we can distinguish between three basic types of (news) recommendation systems: content-based recommenders that recommend news items to users according to the content of the news items, collaborative filtering recommenders that focus on the articles similar users are interested in, and hybrid recommenders that combine the two previous methods. Therefore, the main difference between content-based recommenders and collaborative filtering recommenders lies in the focus on user-article and user-user similarities. In our efforts, the scope will be limited to content-based recommenders. In these recommenders the emphasis is on measuring the similarity between the user profile, built on the content of previously read news articles, and the content of an unread news article. There are two types of content-based recommenders, i.e., traditional and semantic recommenders. The former type is term-based, whereas the latter one is concept-based. The focus in this research will mainly be on semantic recommenders, as these provide for a better and more intuitive representation of news items.

In previous work, Goossen et al. [10] show that using the semantic recommendation method Concept Frequency - Inverse Document Frequency (CF-IDF) significantly improves the performance over a recommender using the traditional Term Frequency - Inverse Document Frequency (TF-IDF). However, a major shortcoming of CF-IDF is that it does not take into account the various semantic relationships between concepts, like the superclass-of, subclass-of, or domain-specific relationships, which could be useful to provide an accurate representation of news items. The goal of this paper is to extend the CF-IDF weighting technique by employing the semantic relationships of concepts from a domain ontology. The proposed recommendation method, CF-IDF+, is implemented in Ceryx [5], an extension to the Hermes news personalization service for building recommender systems [3]. The performance of the CF-IDF+ will be evaluated against the TF-IDF baseline and various semantic-based counterparts like CF-IDF.

The remainder of this paper is organized as follows. First, we discuss related work in Sect. 2. Next, we introduce our framework and implementation in Sects. 3 and 4, respectively. Section 5 presents the performance evaluation of our algorithm, and we conclude in Sect. 6.

2 Related Work

The current body of literature contains many descriptions of profile-based recommender systems, differing in their approaches for news recommendation. Most

importantly, they implement different similarity measures for computing the similarity between a news item and the user profile.

2.1 Term-Based News Recommendation

The most popular traditional content-based recommendation approach is the TF-IDF weighting scheme. A common approach in comparing documents is to use the TF-IDF weighting scheme together with the cosine similarity. It is a statistical method to determine the relative importance for each word within a document in a set of documents.

The TF-IDF weight is defined as a two-step computation, based on the term frequency (TF) and inverse document frequency (IDF). Term frequency indicates the importance of term t_i in document d_j: frequent terms are more likely to be relevant for the topic of the document. The inverse document frequency captures the general importance of a term in a set of documents. The more a term appears in all the documents, the less relevant it is to the topic of a single document, as the term is too generic. The TF-IDF is subsequently computed as the product of TF and IDF.

2.2 Semantics-Based News Recommendation

There are several semantic-based methods available for news recommendation. Some are based on sets of synonyms (synsets), while others are based on ontological concepts.

Synsets. The Synset Frequency - Inverse Document Frequency (SF-IDF) [5] weighting method compares the words from unread documents with the words from documents in the user profile. In the comparison, the semantic similarity is obtained by making use of the WordNet online lexical database. This is a database for the English language with over 166,000 senses (i.e., words with their associated parts-of-speech and senses). The advantage of using synsets over words is that the ambiguity of the words can be eliminated by performing word sense disambiguation.

The SF-IDF weighting method outperforms the traditional TF-IDF, but a shortcoming of this semantics-based method is that they do not allow for various semantic relationships between synsets, thus providing only a limited understanding of news semantics [13]. The SF-IDF recommender does not take into account inter-synset relationships, like hyponymy, antonymy, troponymy, and synonymy, while this is very important for the interpretability of a document.

Moerland et al. [13] extend SF-IDF to the SF-IDF+ weighting method, by accounting for the semantic relationships between synsets. This method is evaluated on a set of financial news articles and it outperforms TF-IDF and SF-IDF in terms of F_1-scores.

Ontological Concepts. Concept Frequency - Inverse Document Frequency (CF-IDF) is based on the traditional TF-IDF method, but merely considers key concepts instead of using all the terms within a document [10]. The TF-IDF recommendation method considers all terms in an article to be useful for understanding the content, while this might lead to 'noise' terms in the weighted vector of terms. These are terms that do not give additional information about the content of the article. Instead of a weighted vector of terms, CF-IDF considers the news article as a weighted vector of concepts, where concepts are derived from a domain ontology. Using concepts instead of terms assures that there are no noisy terms which can obfuscate the algorithm's outcomes, making CF-IDF a more intelligent recommender.

3 Framework

The Hermes news personalization framework introduced in [3] is a news personalization service supporting different content-based recommenders. Hermes operates on RSS feeds of user-rated news items. Recommendations are catered using terms, synsets, or concepts from an internal knowledge base for classification.

3.1 Hermes Extensions

The Hermes News Portal has two extensions, i.e., Athena [11] and Ceryx [5]. Athena supports both the classic TF-IDF and the semantics-based recommendation method CF-IDF, assigning words in news articles to different domain concepts, which are stored in an internal knowledge base. A graph-based user interface allows for visualizing inter-concept connections, and can also be used by users to select interesting concepts to manually build a profile. Automatic user profile construction is also supported by analyzing the articles that are read by the user. The Ceryx extension of Athena provides additional support for synset-based recommenders, such as SF-IDF. In our recent efforts, the newly proposed CF-IDF+ is added to the list, which additionally utilizes semantically related concepts and optimizes concept relationship type weights for enhanced performance.

3.2 User Profile Representation

A user profile is represented by a content vector (with terms, concepts, or synsets) of all user-read news articles. The contents are determined differently depending on the recommender. While the TF-IDF recommender merely analyzes every term in a news item after initial stop word removal, the CF-IDF and CF-IDF+ recommenders look at domain concepts found in the news items instead of using all the terms in the text of the news items.

The Hermes framework employs the vector space model, which is interpreted differently by each recommender. Concept-based recommenders use a domain

ontology to retrieve the concepts in a news item. This ontology holds a set C of i concepts and their relations, i.e., $C = \{c_1, c_2, c_3, \ldots, c_i\}$. The user profile U consist of j concepts, where $U = \left\{c_1^u, c_2^u, c_3^u, \ldots, c_j^u\right\}$ and $c^u \in C$. Concept c^u is linked to k news articles a_i in which it is found, and is denoted as $c^u = \{a_1^u, a_2^u, a_3^u, \ldots, a_k^u\}$. Hence, for a semantic recommender, an article a is considered to be a set of l elements representing the number of concepts c present, i.e., for CF-IDF:

$$a_{\text{CF-IDF}} = \{c_1^a, c_2^a, c_3^a, \ldots, c_l^a\}, c^a \in C \ . \tag{1}$$

The traditional TF-IDF recommender has a different interpretation, as it regards article a as a set of m elements representing the number of terms t appearing in article a:

$$a_{\text{TF-IDF}} = \{t_1^a, t_2^a, t_3^a, \ldots, t_m^a\} \ . \tag{2}$$

Then, we can consider weights w^u and w^a for the user profile and article, respectively, as item scores. Computational details for the weight of a single item (i.e., term or concept) in a specific article for TF-IDF, CF-IDF, and CF-IDF+ can be found in Sect. 3.3. Together, terms and concepts and their respective computed weights comprise the user profile U as a vector V^U, and an unread news item a as vector V^a.

The TF-IDF recommender computes a weight w^u for each term in U and the CF-IDF recommender computes a weight w^u for every concept in U:

$$V^U_{\text{TF-IDF}} = \{\langle\ t_1^u, w_1^u\rangle\ , \ldots, \langle\ t_m^u, w_m^u\rangle\ \}\ , \tag{3}$$

$$V^U_{\text{CF-IDF}} = \left\{\langle\ c_1^u, w_1^u\rangle\ , \ldots, \langle\ c_j^u, w_j^u\rangle\ \right\}\ . \tag{4}$$

When computing recommendations, the TF-IDF and CF-IDF recommender convert unread news item a into vector $V^a_{\text{CF}-\text{IDF}}$, containing the terms or concepts found in the news items and their corresponding weights, respectively:

$$V^a_{\text{TF-IDF}} = \{\langle\ c_1^a, w_1^a\rangle\ , \ldots, \langle\ c_m^a, w_m^a\rangle\ \}\ , \tag{5}$$

$$V^a_{\text{CF-IDF}} = \{\langle\ c_1^a, w_1^a\rangle\ , \ldots, \langle\ c_l^a, w_l^a\rangle\ \}\ . \tag{6}$$

A similarity measure, such as the cosine similarity, can subsequently be applied to the user profile and news item vectors, resulting in a ranked recommendation based on similarity.

3.3 Weight Computation

In order to ensure a general understanding of the computation of CF-IDF+ weights, we briefly discuss TF-IDF and CF-IDF weight computations. In TF-IDF computations, we are interested in the product of term frequencies tf and inverse document frequencies idf. The term frequency is defined as the number of times a term t_i occurs in document d_j, $n_{i,j}$, divided by the total number of occurrences of all terms in the document, whereas the inverse document frequency indicates

the importance of term t in all news items and is computed by dividing the total number of documents $|D|$ by the number of documents in which term t_i can be found:

$$\text{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \ , \tag{7}$$

$$\text{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \ , \tag{8}$$

$$\text{tf-idf}_{i,j} = \text{tf}_{i,j} \cdot \text{idf}_i \ . \tag{9}$$

The similarity between the user profile vector and the vector representation of a news item is subsequently calculated using the cosine similarity measure, i.e.:

$$\text{sim}_{\text{TF-IDF}} = \frac{P \cdot U}{\|P\| \|U\|} \ , \tag{10}$$

where the user profile vector is represented by P, and U is the vector representation with the terms in the unread news item. If this similarity score is greater than the cut-off value, the unread news item is recommended to the user. The cut-off value can be any number between 0 and 1. Lower cut-off values will result in more recommended news items (high recall) with a higher likelihood of them being non-interesting (low precision), whereas higher values enforce high similarities and thus yield higher precision and lower recall. Therefore, the cut-off value is an indication of the degree of the user's preference.

The procedures of the Concept Frequency - Inverse Document Frequency recommender [10] are very similar to the TF-IDF recommender, yet operate on concept vectors instead of term vectors. Each news item contains zero or more concepts, which are defined in the domain ontology of Hermes. To obtain these concepts, Hermes employs a Natural Language Processing engine that uses several techniques like tokenization, part-of-speech tagging and word sense disambiguation. Similar to TF-IDF frequencies and inverse document frequencies are computed for concept c_i, i.e., cf and idf$'$, respectively:

$$\text{cf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \ , \tag{11}$$

$$\text{idf}'_i = \log \frac{|D|}{|\{d : c_i \in d\}|} \ , \tag{12}$$

$$\text{cf-idf}_{i,j} = \text{cf}_{i,j} \cdot \text{idf}'_i \ . \tag{13}$$

The similarity between the user profile vector and the vector representation of the unread news item is computed by the cosine similarity measure introduced before. Again the unread news article U is only recommended if the similarity measure exceeds the cut-off value.

The usage of concepts instead of terms causes the recommender to deal with only the most 'important' terms in an article, making it more effective. Even when considering the additional computation time caused by necessary NLP steps, processing times are much lower for CF-IDF when compared to TF-IDF, because the number of elements that need to be processed is much lower.

Despite its benefits, an earlier mentioned drawback of CF-IDF is that it does not take into account concept relationship types. These types provide valuable

information about the content of a news item, that can thus be exploited to retrieve interesting news items that discuss related concepts. The CF-IDF+ recommender also processes the news items into a concept vector representation, but additionally verifies whether the concept is a class or an instance. For a class, its direct super- and subclasses are retrieved from the domain ontology. For an instance, its domain relationships are identified from the domain ontology.

For our subsequent procedures, we extend the original set of concepts C to also include the concepts that are related through various semantical relationships of the included concepts, and define $C(c) = \{c\} \bigcup_{r \in R(c)} r(c)$, where c denotes a concept in the news item, $r(c)$ denotes a concept that is related to concept c by relationship r and $R(c)$ represents the set of relationships of concept c. Next, we define two sets of extended concepts, i.e., the unread news item U and the user profile R as $U = \{C(u_1), C(u_2), \ldots, C(u_m)\}$, and $R = \{C(r_1), C(r_2), \ldots, C(r_n)\}$, respectively, where $C(u_m)$ denotes the m-th extended concept of U, and $C(r_n)$ represents the n-th extended concept of R. By extended concept we mean all the concepts obtained by following relationships in the domain ontology from the current concept.

The calculation of the CF-IDF+ weight is very similar to the original CF-IDF, but additionally introduces weights per semantic relationship. This has also proven to be very effective in a word sense disambiguation context, by weighting edges (relations) of network representations of semantic lexicons [15]. In CF-IDF+, we focus on domain ontologies, and identify three different weights: one for the superclasses, one for the subclasses and one for the domain relationships. CF-IDF+ is computed as follows:

$$\text{cf-idf}_{i,j,r} = \text{cf}_{i,j} \cdot \text{idf}'_i \cdot w_r \ . \tag{14}$$

The weight of semantic relationship r needs to be optimized for each cut-off value, for instance through genetic algorithms or brute-force iterative processes (e.g., explicit enumeration). If a concept is found multiple times (direct occurrence in text and indirect occurrence by means of a relationship), we retain only its highest weight value.

4 Implementation

In our endeavours, we have implemented CF-IDF+ in Ceryx, which is an extension to the Java-based Hermes News Portal and its Athena plug-in for recommender systems. The tool makes use of various Semantic Web technologies, operates on user profiles, and processes news items from RSS feeds. The core of the news portal is an OWL domain ontology that is constructed by domain experts, allowing for semantics-based operations on news messages. The ontology contains a set of commonly used, well-known, financial entities such as companies, products, currencies, etc., and these concepts have associated lexical representations. The ontology consists of 65 classes, 18 object properties, 11 data properties, and 1,167 individuals.

News items are classified using the GATE natural language processing software [8] and the WordNet [9] semantic lexicon. The semantics-based methods additionally make use of the Stanford Log-Linear Part-of-Speech Tagger [16], Lesk Word Sense Disambiguation [12], and the Alias-i's LingPipe 4.1.0 [1] Named Entity Recognizer.

4.1 User Interface

The Ceryx implementation features a tabbed interface in which the user is able to browse through all available news items in the system, which are extracted from RSS feeds. Whenever an item is clicked, the corresponding news message is opened in the Web browser and the user profile is updated. Additionally, the user is able to select a recommendation method. Retrieved recommendations are sorted by relevance to the user's profile. Last, we have implemented a testing environment that compares the various news recommenders implemented in Ceryx. This environment supports XML input files describing user profiles (i.e., lists with the URIs of the read news items and whether the news items are considered as interesting by the user), and displays common evaluation measures, such as precision and recall, based on randomly selected training and test sets. Cut-off testing is also supported, and is preliminary implemented as an iterative process with increments of 0.01 from 0 to 1.

4.2 TF-IDF

For each news item, the TF-IDF news recommender gathers all words (terms), and removes stop words using a predefined list. Subsequently, term counts are computed within each news item and over all news items, which are used for computing term frequencies and inverse document frequencies. Before the news items are recommended to the user, the similarity scores for all the unread news items are MIN-MAX normalized to take values between 0 and 1. If the similarity score exceeds the user-specified cut-off value, the unread news item will be recommended.

4.3 CF-IDF

In order to retrieve concepts from the news items, Ceryx utilizes an advanced Natural Language Processing engine that matches the lexical representations stored in the ontology and subsequently performs word sense disambiguation. When the concepts are obtained from the news items, the CF-IDF recommender operates similarly to the TF-IDF recommender. The CF-IDF recommender counts the number of concept appearances in each news items to obtain the concept frequencies, as well as the number of times a concept is found in all news items to collect the inverse document frequencies. Again, CF-IDF weights are computed and similarities are MIN-MAX normalized. As before, if the similarity score exceeds the user-specified cut-off value, the unread news item will be recommended.

4.4 CF-IDF+

The implementation of CF-IDF+ is an extension to the CF-IDF implementation and additionally determines whether identified concepts are classes or instances, so that related concepts are retrieved. The CF-IDF values of these related concepts are calculated by multiplying the CF-IDF of the original concept with a weight depending on the relationship type between the original concept and the related concept. The CF-IDF+ values are then stored in a vector for the news item or the user profile. When a related concept is already part of the vector, we verify whether the CF-IDF+ value of the related concept is greater than the original CF-IDF+ of the related concept, and choose the highest CF-IDF+ value. Next, the similarity score between the user profile and each news item is computed by using the cosine similarity measure. Again, similarity scores are MIN-MAX normalized between 0 and 1, and if the similarity score exceeds the user-specified cut-off value, the unread news item will be recommended.

5 Evaluation

In order to demonstrate the effectiveness of CF-IDF+, we evaluate the performance of its implementation within the Ceryx plugin of the Hermes news recommendation system against real-world data. The remainder of this section discusses the experimental setup, the optimization of the semantic relationship weights and their properties, and the evaluation of the global performance of CF-IDF+ compared to various other recommenders, respectively.

5.1 Data

In our experiments, we make use of a data set containing 100 real-life news articles, which are collected from a Reuters news feed on news about technology companies. We distinguish between 8 topics of interest and ask 3 domain experts to rate each news item for appropriateness with respect to the selected topics, as described in Table 1, i.e., 'Asia or its countries', 'Financial markets', 'Google or its competitors', 'Internet or Web services', 'Microsoft or its competitors', 'National economies', 'Technology', and 'United States'. We maintain an Inter-Annotator Agreement (IAA) of at least 2 out of 3 reviewers in order for a news item to be added to a user profile for one of the specific subjects. Table 1 described the number of interesting (I+) and non-interesting (I-) items per topic, and additionally shows the fraction of interesting news items indicated as interesting by all domain experts (IAA+), the fraction of non-interesting news items indicated as non-interesting by all all experts (IAA-) and the average of these two fractions (IAA).

For evaluation purposes, we split our data randomly into a 30% training, a 30% validation, and a 40% test set. These sets are used for creating user profiles, optimizing relationship type weights, and performance measurement, respectively. Based on true/false positives/negatives, we evaluate performance

Table 1. Overview of topics and their associated number of (non-)interesting news items, accompanied by their inter-annotator agreements.

Topic	I+	I-	IAA+	IAA-	IAA
Asia or its countries	21	79	100%	98%	99%
Financial markets	24	76	75%	68%	72%
Google and its competitors	26	74	100%	95%	97%
Internet or Web services	26	74	96%	92%	94%
Microsoft or its competitors	29	71	100%	96%	98%
National economies	33	67	94%	85%	90%
Technology	29	71	86%	87%	87%
United States	45	55	87%	84%	85%
Average	29	71	92%	88%	90%

for various recommendation cut-off values by computing accuracy, recall, precision, and F_1 scores. Weights are optimized using an iterative procedure with a step size of 0.1 between 0 and 1, while aiming to maximize F_1 scores.

5.2 Semantic Relationships Weights

Per cut-off value, ranging between 0 and 1 with an increment of 0.01, we optimize the weights of the superclass, subclass, and domain semantic relationships. The mean and variance of these weights (w_{super}, w_{sub}, and w_{rel}, respectively) are displayed in Table 2. Generally, concepts retrieved through domain relationships seem to be of more importance than sub- and superclasses, and concepts retrieved through superclasses are of more importance than concepts retrieved through subclasses. Subclasses have a tendency to be too specific, while superclasses give more general information. Domain relationships on the other hand define properties of the original concept, and are hence more likely to be more valuable for extending concepts. For example, the concept 'Windows', an instance in the ontology, has the domain relationships `hasUpdate` 'Fall Creators Update' and `isProducedBy` 'MSFT'. These related concept are closely linked to the original concept 'Windows'. Its superclass 'Operating System' is too generic, and its subclass 'Windows Mobile' is too specific.

5.3 Experimental Results

Next, we evaluate the experimental results based on the optimized weights for each cut-off value. The precision, recall, and F_1 scores for the CF-IDF+, CF-IDF,

Table 2. Mean and variance of the weights for the semantic relationships.

	w_{super}	w_{sub}	w_{rel}
μ	0.330693	0.161386	0.500000
σ^2	0.126149	0.069594	0.147400

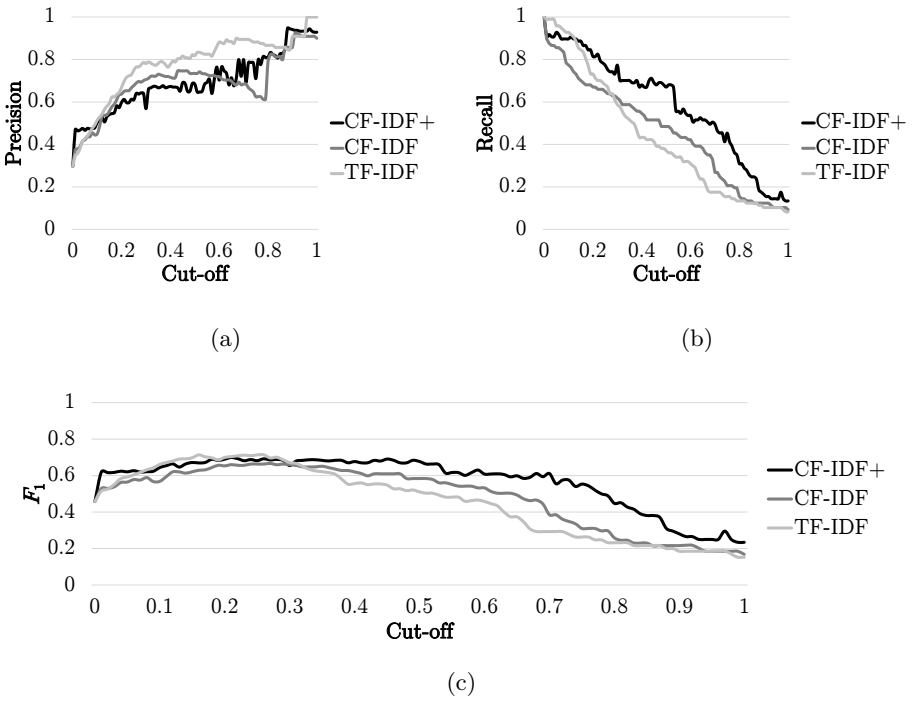

Fig. 1. Global precision, recall, and F_1 scores for the CF-IDF+, CF-IDF, and TF-IDF recommenders for varying cut-off values.

and TF-IDF recommenders are plotted against the cut-off values in Figs. 1a, 1b, and 1c. In terms of F_1, overall, the CF-IDF+ outperforms the original CF-IDF method, although for the cut-off values between 0 and 0.3, the difference between CF-IDF+ and CF-IDF is rather small, and at times, both concept-based recommenders are outperformed by TF-IDF.

The performance of the CF-IDF+ recommender is much more stable over all cut-off values, than the other two recommenders. The high F_1 scores are a result of a much higher recall, as CF-IDF(+) precision is lower than TF-IDF precision. CF-IDF+ clearly outperforms the TF-IDF and CF-IDF recommenders on the recall measure, except for very low cut-off values, when TF-IDF obtains a higher recall. The CF-IDF+ method has a lower precision than the TF-IDF recommender and a relatively similar precision to the CF-IDF recommender. Despite the, at times, disappointing precision scores, the potential benefits of CF-IDF+ remain evident. For low cut-off values, the CF-IDF+ recommender has a higher precision than TF-IDF due to the inherent noise removal by considering concepts rather than terms. The limited ontology quality (i.e., the coverage of concepts and their lexical representations) comes into play for higher cut-off values. Naturally, CF-IDF+'s lower precision could be remedied by improving

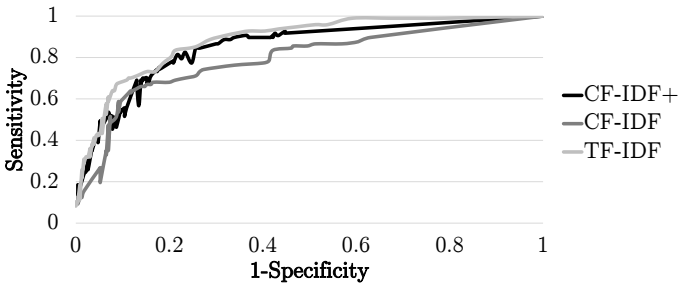

Fig. 2. ROC curves for the CF-IDF+, CF-IDF, and TF-IDF recommenders.

the employed natural language processing pipeline, more precisely the word sense disambiguation (currently based on the rather basic adapted Lesk algorithm [2]).

Next, we evaluate the Receiver Operating Characteristic (ROC) curves for the CF-IDF+, CF-IDF, and TF-IDF recommenders, which plot recall against the percentage of uninteresting rated news items defined as interesting by the recommender, i.e., the False Positive Rate (FPR) that is computed as $1-$specificity (fall-out), while varying the cut-off value. The ROC curves in Fig. 2 show that the CF-IDF+ generally outperforms the CF-IDF recommender, and performs comparably to the TF-IDF recommender. The performances are also underlined by the Area Under the Curve (AUC). For the ROC curves in Figure 2, the AUC of the CF-IDF+, CF-IDF, and TF-IDF recommenders are 0.85072, 0.79162, and 0.88274, respectively.

The Kappa statistic [7] measures whether the proposed classifications made by the recommender are better than random guessing. The closer the statistic is to 1, the more classification power a recommender has. Figure 3 shows that the

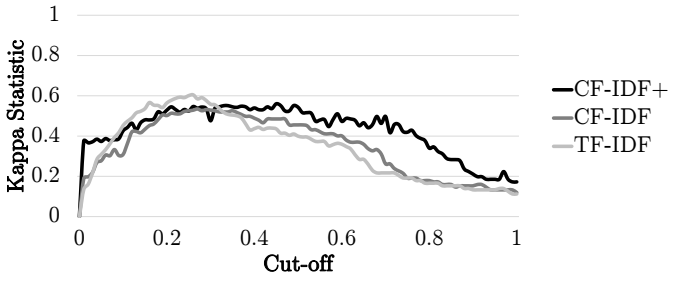

Fig. 3. Kappa statistics for the CF-IDF+, CF-IDF, and TF-IDF recommenders for varying cut-off values.

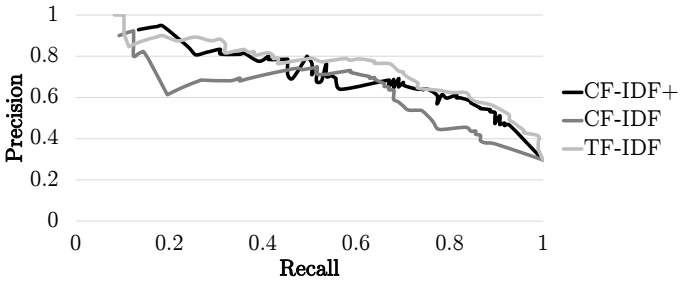

Fig. 4. PR curves for the CF-IDF+, CF-IDF, and TF-IDF recommenders.

Kappa statistic for CF-IDF+ is significantly higher than for CF-IDF and TF-IDF. For the cut-off values between 0.08 and 0.30, the TF-IDF performs slightly better than the CF-IDF+ and CF-IDF recommenders, but overall, the CF-IDF+ has more classification power than the CF-IDF and TF-IDF recommenders.

The last curve that helps us to assess the performance of our CF-IDF+ recommender is the Precision-Recall (PR) curve. In general, a higher precision is associated with a lower recall, and vice versa. Figure 4 demonstrates that for the 3 recommenders, this also holds. However, even though TF-IDF was outperformed based on F_1, this is clearly not the case judging by the PR curves. The TF-IDF recommender outperforms both the CF-IDF+ and CF-IDF recommenders. The higher F_1 scores can be explained by the fact that most points of CF-IDF+ and CF-IDF are centered around the middle of the graph, resulting in higher F_1 scores due to the characteristics of the F_1 calculation. TF-IDF produces a more evenly distributed set of precision and associated recall values. Overall, CF-IDF+ still performs notably better than the CF-IDF recommender. For example, for a recall of approximately 0.2, the CF-IDF method has a precision of about 0.6, while the CF-IDF+ has a precision higher than 0.9. Therefore, the CF-IDF+ recommender balances precision and recall in a better way than the CF-IDF recommender.

6 Conclusions

In our endeavours, we have aimed to improve the semantics-based CF-IDF (news) recommendation method that operates on ontologies, by additionally taking into account semantic inter-concept relationships, hereby extending the search space. For ontological classes, direct sub- and superclasses are included in the similarity analyses, and for ontological instances, concepts which are related by domain relationships are also taken into consideration. Additionally, we have optimized the semantic relationship weights based on the global F_1-scores.

When comparing our CF-IDF+ recommender to the original CF-IDF and baseline TF-IDF recommenders, we learned from the ROC curve, the Kappa

statistic, and from precision, recall, and F_1-scores, that the CF-IDF+ shows a significant improvement over the original CF-IDF recommender. The CF-IDF+ recommender also performs better than the TF-IDF recommender in terms of recall, F_1, and the Kappa statistic. However, the TF-IDF recommender did have a better balance between fall-out and sensitivity, and precision and recall.

We envision various directions for future work. First, a more fine-grained (learning) approach to semantic relationship weight optimization would provide additional insights and could possibly enhance CF-IDF+ performance. Furthermore, we would like to investigate a larger collection of relationships. Now, we have considered the direct super- and subclasses, but hypothetically, non-direct super- and subclasses of concepts could be valuable as well. Last, a more thorough and powerful evaluation based on a larger set of news items would further underline the strong performance of CF-IDF+.

Acknowledgement

The authors would like to thank Tim Vos for his support and the fruitful discussions on this topic.

References

1. Alias-i: LingPipe 4.1.0. From: `http://alias-i.com/lingpipe` (2017)
2. Banerjee, S., Pedersen, T.: An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: Gelbukh, A.F. (ed.) 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2002). pp. 136–145. Springer-Verlag (2002)
3. Borsje, J., Levering, L., Frasincar, F.: Hermes: a Semantic Web-Based News Decision Support System. In: 23rd Annual ACM Symposium on Applied Computing (SAC 2008). pp. 2415–2420. ACM (2008)
4. Capelle, M., Hogenboom, F., Hogenboom, A., Frasincar, F.: Semantic News Recommendation Using WordNet and Bing Similarities. In: Shin, S.Y., Maldonado, J.C. (eds.) 28th Symposium on Applied Computing (SAC 2013), The Semantic Web and its Application Track. pp. 296–302. ACM (2013)
5. Capelle, M., Moerland, M., Frasincar, F., Hogenboom, F.: Semantics-Based News Recommendation. In: Akerkar, R., Bădică, C., Dan Burdescu, D. (eds.) 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS 2012). ACM (2012)
6. Capelle, M., Moerland, M., Hogenboom, F., Frasincar, F., Vandic, D.: Bing-SF-IDF+: A Hybrid Semantics-Driven News Recommender. In: Wainwright, R.L., Corchado, J.M., Bechini, A., Hong, J. (eds.) 30th Symposium on Applied Computing (SAC 2015), Web Technologies Track. pp. 732–739. ACM (2015)
7. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
8. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002). pp. 168–175. Association for Computational Linguistics (2002)

9. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
10. Goossen, F., IJntema, W., Frasincar, F., Hogenboom, F., Kaymak, U.: News Personalization using the CF-IDF Semantic Recommender. In: Akerkar, R. (ed.) International Conference on Web Intelligence, Mining and Semantics (WIMS 2011). ACM (2011)
11. IJntema, W., Goossen, F., Frasincar, F., Hogenboom, F.: Ontology-Based News Recommendation. In: Daniel, F., Delcambre, L.M.L., Fotouhi, F., Garrigós, I., Guerrini, G., Mazón, J.N., Mesiti, M., Müller-Feuerstein, S., Trujillo, J., Truta, T.M., Volz, B., Waller, E., Xiong, L., Zimányi, E. (eds.) International Workshop on Business intelligencE and the WEB (BEWEB 2010) at 13th International Conference on Extending Database Technology and 13th International Conference on Database Theory (EDBT/ICDT 2010). ACM (2010)
12. Jensen, A.S., Boss, N.S.: Textual Similarity: Comparing Texts in Order to Discover How Closely They Discuss the Same Topics. Bachelor's Thesis, Technical University of Denmark (2008)
13. Moerland, M., Hogenboom, F., Capelle, M., Frasincar, F.: Semantics-Based News Recommendation with SF-IDF+. In: Camacho, D., Akerkar, R., Rodríguez-Moreno, M.D. (eds.) 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS 2013). ACM (2013)
14. Ostuni, V.C., Noia, T.D., Sciascio, E.D., Mirizzi, R.: Top-N Recommendations from Implicit Feedback Leveraging Linked Open Data. In: 7th ACM Conference on Recommender Systems (RecSys 2013). pp. 85–92. ACM (2013)
15. Sussna, M.: Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network. In: Bhargava, B., Finin, T., Yesha, Y. (eds.) 2nd International Conference on Information and Knowledge Management (CIKM 1993). pp. 67–74. ACM (1993)
16. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLTNAACL 2003). pp. 252–259 (2003)