Semantics-Based News Recommendation with SF-IDF+

Marnix Moerland
marnix.moerland@gmail.com

Michel Capelle
michelcapelle@gmail.com

Frederik Hogenboom
fhogenboom@ese.eur.nl

Flavius Frasincar
frasincar@ese.eur.nl

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

ABSTRACT

Content-based news recommendations are usually made by employing the cosine similarity and the TF-IDF weighting scheme for terms occurring in news messages and user profiles. Recent developments, such as SF-IDF, have elevated news recommendation to a new level of abstraction by additionally taking into account term meaning through the exploitation of synsets from semantic lexicons and the cosine similarity. Other state-of-the-art semantic recommenders, like SS, make use of semantic lexicon-driven similarities. A shortcoming of current semantic recommenders is that they do not take into account the various semantic relationships between synsets, providing only for a limited understanding of news semantics. Therefore, we extend the SF-IDF weighting technique by additionally considering the synset semantic relationships from a semantic lexicon. The proposed recommendation method, SF-IDF+, as well as SF-IDF and several semantic similarity lexicon-driven methods have been implemented in Ceryx, an extension to the Hermes news personalization service. An evaluation on a data set containing financial news messages shows that overall (by accounting for all considered cut-off values) SF-IDF+ outperforms TF-IDF, SS, and SF-IDF in terms of F_1-scores.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Relevance feedback*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Representation Languages*

General Terms

Algorithms, Performance

WIMS'13, June 12–14, 2013, Madrid, Spain.

Keywords

Content-based recommender, News personalization, Recommender systems, Semantic Web, User profiling, Semantic similarity

1. INTRODUCTION

The Web is becoming an increasingly important source of information to many day-to-day users, which is disseminated in the form of news messages, videos, etc. However, users are generally overwhelmed by the amount of information, and hence content-based recommendation methods enjoy an increased attention. These methods filter and structure information and distinguish between interesting and non-interesting news articles, videos, products, etc. Based on user preferences captured in user profiles (usually derived from user browsing behavior), recommendations can be made by comparing new items with a user profile.

Traditionally, content-based recommender systems operate based on term frequencies. A commonly used measure is Term Frequency – Inverse Document Frequency (TF-IDF) [21]. When employing user profiles that describe users' interest based on previously browsed items, these can be translated into vectors of weights, which can be utilized for calculating the interestingness of a new item using a measure like cosine similarity. However, such systems do not take into consideration the text semantics. One could add semantics to such an approach by employing Web ontologies, yet these are domain dependent, requiring continuous maintenance. One could also employ general semantic lexicons such as WordNet [9] and use their synonym sets (synsets), which are in essence collections of words that have associated morphological and semantic information.

Hence, in a previous contribution [6], we introduced the Synset Frequency – Inverse Document Frequency (SF-IDF) measure, operating on WordNet synsets. When evaluating the performance of SF-IDF compared to TF-IDF and Semantic Similarity (SS) [6], we have shown its superiority over the other approaches. However, we did not take into account inter-synset relationships, while research has shown that relationships such as synonymy, hyponymy, merynomy, troponymy, antonymy, and entailment provide more structure in a text and hence contribute to an improved level of interpretability [12]. Therefore, in our current endeavours we extend the SF-IDF weighting technique by additionally considering synset semantic relationships. Together with other state-of-the-art semantic recommenders like SF-IDF and SS, the proposed recommendation method, SF-IDF+, is implemented in Ceryx, an extension to the Hermes news personalization service [4, 10, 11, 13]. Moreover, we evaluate the performance of SF-IDF+ against the TF-IDF baseline and semantics-based approaches as SF-IDF and SS.

The remainder of this paper is organized as follows. First, Section 2 discusses related work. Then, we present the new recommender SF-IDF+ and its implementation in Sections 3 and 4, respectively. Next, we evaluate the results obtained using the new recommender as well as existing ones on a news corpus in Section 5. Last, we conclude our paper and propose future work in Section 6.

2. RELATED WORK

Over the years, many recommender systems have been developed that each make use of user profiles, yet they differ in their approaches for news recommendation. This section continues by elaborating on existing recommender systems, and subsequently discusses various content-based recommendation methods such as TF-IDF, SF-IDF, and SS. Last, we describe related work on semantic relations and their previous exploitation in news processing.

2.1 Recommender Systems

The NewsDude [3], NewsWeeder [16], and Webclipping [7] recommender systems construct user profiles through user-guided elicitation interfaces. Subsequently, NewsWeeder employs an algorithm based on the discriminatory power of words in order to predict ratings for unseen news items. NewsDude on the other hand employs Term Frequency – Inverse Document Frequency (TF-IDF) [21] vectors for modeling the user's short-term interest, and Boolean feature vectors for modeling the user's long-term interests. Then, a model that uses a combination of the Nearest Neighbour algorithm and the cosine similarity proposes articles for short-term interest, whereas long-term interest articles are suggested by a model that makes use of the Boolean feature vectors and a naïve Bayesian classifier. Webclipping follows a different strategy, as it assigns values indicating the likeliness of appearing in the user profile to keywords from an unread news item by means of a Bayesian classifier, after which Bayes' rule is applied to rate the overall similarity of an unread news item to the user profile.

The Krakatoa Chronicle [2] is a hybrid recommendation system which combines content-based filtering and collaborative filtering. As most other recommenders, the user profile consists of a vector of keywords of the articles which the user has read and deemed interesting. However, the framework distinguishes between personal interestingness and general interestingness, and hence makes use of several user profiles at the same time. Last, the News@Hand [5] recommender system uses semantic information from Wikipedia. Terms are matched to ontology classes and news items are matched to the user profile through cosine-based vector similarities.

2.2 TF-IDF

One of the most commonly used similarity measures is TF-IDF [21], especially in combination with cosine similarities. The method operates on terms T in documents D and consists of a term frequency $tf(t,d)$ and an inverse document frequency $idf(t,d)$. The term frequency $tf(t,d)$ measures the number of occurrences n of term $t \in T$ in document $d \in D$ expressed as a fraction of the total number of occurrences of all k terms in document d:

$$tf(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \ . \tag{1}$$

The inverse document frequency on the other hand expresses the number of occurrences of a term t in a set of documents D:

$$idf(t,d) = \log \frac{|D|}{|\{d : t \in d\}|} \ . \tag{2}$$

Here, $|D|$ is the total number of documents in the set of documents D that are compared. The amount of documents which contain term t is denoted by $d : t \in d$. Next, TF-IDF values are calculated through the multiplication of $tf(t,d)$ and $idf(t,d)$:

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t,d) \ . \tag{3}$$

For all terms t in all documents d, TF-IDF value are stored in a vector $A(d)$, after which the similarity between a set of terms from news item d_u and user profile d_r is calculated using the cosine similarity measure, which is defined as:

$$sim_{tf\text{-}idf}(d_u, d_r) = \frac{A(d_u) \cdot A(d_r)}{||A(d_u)|| \times ||A(d_r)||} \ . \tag{4}$$

After every unread document has been assigned a value representing its similarity with the user profile, the unread news items with a similarity value higher than a cut-off value are recommended to the user.

2.3 SF-IDF

As mentioned before, an important drawback of TF-IDF-based recommendation is that semantics are not taken into account. Therefore, different words with the same meaning would be counted as two separate terms, and a word appearing for two different meanings will be counted as one. Because of this, in earlier work, the Synset Frequency – Inverse Document Frequency (SF-IDF) [6] has been proposed, which is a TF-IDF-based measure that additionally makes use of synonym sets (synsets) from a semantic lexicon such as WordNet [9]. These synsets are obtained after performing word sense disambiguation using the adapted Lesk algorithm [1]. After replacing term t by synset s, the SF-IDF formulas are:

$$\begin{aligned} sf\text{-}idf(s,d) =& sf(s,d) \times idf(s,d) \ , & (5)\\ sim_{sf\text{-}idf}(d_u, d_r) =& \frac{A(d_u) \cdot A(d_r)}{||A(d_u)|| \times ||A(d_r)||} \ . & (6) \end{aligned}$$

Similar to the TF-IDF method, the cosine similarity measure is used for computing similarity scores of unread news articles with respect to the user profile. Unread news items that have obtained a rating above a specific cut-off value are suggested to the user.

2.4 Semantic Similarity

Another semantics-based measure introduced in related work is the Semantic Similarity (SS) [6]. The synsets are computed in a similar way as in the previously presented method. Here, synsets from unread news items are compared with user profile synsets. Let us consider vector V, containing all possible combinations of synsets from unread news item d_u, U, and the union of synsets from the user profile d_r, R:

$$V = (\langle u_1, r_1 \rangle, \ldots, \langle u_k, r_l \rangle) \; \forall \; u \in U, \; r \in R \; . \tag{7}$$

Here, u_k denotes a synset from the unread news item, r_l is a synset from the user profile, and k and l represent the number of synsets in the unread news item and the user profile, respectively. Now, let W be a subset of V, containing all the combinations that have a common part-of-speech:

$$W \subseteq V \; \forall \; (u, r) \in W : POS(u) = POS(r) \; . \tag{8}$$

Here, $POS(u)$ and $POS(r)$ are defined as the part-of-speech of synsets u and r in the unread news item and user profile, respectively.

For every combination in W, a similarity rank is computed, measuring the semantic distance between synsets u and r when represented as nodes in a hierarchy of 'is-a' relationships:

$$sim_{SS}(W) = \frac{\sum\limits_{(u,r) \in W} sim(u, r)}{|W|} \; , \tag{9}$$

where $sim(u, r)$ is the similarity rank between the synsets u and r, and $|W|$ denotes the number of combinations between the synsets from the unread news item and the user profile. Again, the unread news which rank higher than a specific cut-off value are recommended to the user.

Existing literature reports on various similarity measures, some of which are based on the information content IC of the synset nodes:

$$IC(s) = - \log \sum_{w \in s} p(w) \; , \tag{10}$$

where $p(x)$ denotes the probability that an instance x of synset s occurs in a corpus, and w represents a word in synset s. The Jiang & Conrath [15] measure $sim_{J\&C}$ makes use of the information content of both the synsets and of the lowest common subsumer (LCS), whereas Lin's similarity measure [19] sim_L uses the logarithms of the chances of appearance of both synsets and the lowest common subsumer. Last, Resnik's measure [20] sim_R maximizes the information content of the lowest common subsumer of the two nodes. The previous information content-based measures are defined as follows:

$$sim_{J\&C}(u, r) = \frac{1}{IC(u) + IC(r) - 2 \times IC(LCS(u, r))} \; , \tag{11}$$

$$sim_L(u, r) = \frac{2 \times \log p(LCS(u, r))}{\log p(u) + \log p(r)} \; , \tag{12}$$

$$sim_R(u, r) = IC(LCS(u, r)) \; . \tag{13}$$

Other similarity measures make use of path lengths between nodes, e.g., the Leacock & Chodorow [17] and Wu & Palmer [24] measures. The path length is either the shortest path (Λ) between the two nodes or the depth (Ω) from a node to the top node. Leacock & Chodorow's measure $sim_{L\&C}$ makes use of the shortest path length Λ between nodes u and r, while Wu & Palmer's similarity measure $sim_{W\&P}$ is based on the depth of the lowest common subsumer of both nodes and the shortest path length between them:

$$sim_{L\&C}(u, r) = - \log \frac{\Lambda(u, r)}{2\Omega} \; , \tag{14}$$

$$sim_{W\&P}(u, r) = \frac{2 \times \Omega(LCS(u, r))}{\Lambda(u, r) + 2 \times \Omega(LCS(u, r))} \; . \tag{15}$$

2.5 Semantic Relations

Even though the SF-IDF and SS methods incorporate text semantics, they do not take into account semantic relations. Research has shown that inter-concept relationships provide more structure to a text, hereby contributing to text interpretability [12]. The authors of [12] propose the usage of Semantic Relatedness for news similarities computation. They make use of WordNet relationships regarding synonymy, hyponymy, and meronomy and weights are assigned by means of maximum enclosure similarities. Getahun et al. however merely employ a limited set of relations, whereas in our work, we utilize all relationships available in WordNet, as we hypothesize that there are more semantic relations that can help reveal text semantics. Moreover, we aim for a more advanced mechanism for determining importance weights for these relationships in the form of a machine learning approach.

3. SF-IDF+ NEWS RECOMMENDATION

SF-IDF+ performs recommendations based on a user profile, which reflects the user's interests. We assume that users only read news items to their likings, and thus construct a user profile by including all currently read news items. A user profile is updated upon reading previously unseen news items by the user. For every unread news item, a similarity score between the news article and the user profile is computed, which is based on similarities between synsets. Unread news items having a similarity scores that exceed a specific cut-off value are recommended to the user.

The SF-IDF+ similarity score takes into account sets of synonyms (synsets) of words stemming from a semantic lexicon such as WordNet [9], and is based on the SF-IDF similarity measure introduced in earlier work [6]. As a preprocessing step, all synsets are retrieved from the unread news items by means of natural language processing techniques [6]. Then, the set of synsets is extended by appending the concepts that are referred to by semantical relationships of the included synsets, and is defined as:

$$S(s) = \{s\} \cup \bigcup_{r \in R(s)} r(s) \; . \tag{16}$$

Here, s denotes the synset in the news item, $r(s)$ represents the synset that is related to synset s by relationship r, and last, $R(s)$ is the set of relationships of synset s from a semantic lexicon.

Next, we define two sets of extended synsets, U and R:

$$U = \{S(u_1), S(u_2), \ldots, S(u_k)\} \; , \tag{17}$$

$$R = \{S(r_1), S(r_2), \ldots, S(r_l)\} \; , \tag{18}$$

which represent the unread news item and the user profile. Here, $S(u_k)$ is the k-th extended synset in the set of extended synsets of the unread news item d_u, U, and $S(r_l)$ denotes the l-th extended synset in the set of extended synsets of the user profile d_r, R.

The computation of SF-IDF+ values is similar to SF-IDF and TF-IDF calculations introduced earlier in Section 2. However, SF-IDF+ uses extended synsets instead of terms (as is the case in the TF-IDF recommendation) or synsets (as is done in the SF-IDF recommendation). Also, weighting is applied depending on the relationship that the semantically related synset has with the synset:

$$sf\text{-}idf{+}(s,d,r) = sf(s,d) \times idf(s,d) \times w_r \; . \quad (19)$$

Here, $d \in \{d_u, d_r\}$ (with d_u and d_r denoting an unread news item and user profile, respectively). Subsequently, $sf(s,d)$ denotes the synset frequency of synset s in the unread news item or the user profile d, $idf(s,d)$ is the inverse document frequency of synset s in d (taken over all news items), and w_r is the weight of the relationship r between the semantically related synset (by means of r) and the synset s. The latter weight can be optimized in a later stage, for example by means of a genetic algorithm.

Subsequently, two vectors are constructed that represent the unread news item d_u and the user profile d_r. These vectors each contain the $sf\text{-}idf(s,d)$ and $sf\text{-}idf{+}(s,d,r)$ values for all (extended) synsets s in d:

$$A(d) = \begin{Bmatrix} \varsigma(s_1,d), \varsigma(s_1,d,r_1), \ldots, \varsigma(s_1,d,r_{m_{s_1}}), \\ \varsigma(s_2,d), \varsigma(s_2,d,r_1), \ldots, \varsigma(s_2,d,r_{m_{s_2}}), \\ \cdots \\ \varsigma(s_n,d), \varsigma(s_n,d,r_1), \ldots, \varsigma(s_n,d,r_{m_{s_n}}) \end{Bmatrix} , \quad (20)$$

where $\varsigma(s,d)$ represents $sf\text{-}idf(s,d)$, and $\varsigma(s,d,r)$ represents $sf\text{-}idf{+}(s,d,r)$. Furthermore, m is the size of the synset set S, n denotes the size of the set of directly found synsets s, and m_{s_n} is the total number of synsets related to synset s_n.

Next, we can compute the similarity score between the unread news item d_u and the user profile d_r by means of the cosine similarity measure, which is defined as:

$$sim_{sf\text{-}idf+}(d_u,d_r) = \frac{A(d_u) \cdot A(d_r)}{||A(d_u)|| \times ||A(d_r)||} \; . \quad (21)$$

4. SF-IDF+ IMPLEMENTATION

Our semantics-based recommendation method is implemented as an extension to the Ceryx [6] plugin of the Hermes News Portal (HNP) [4, 10, 11, 13], which is a news recommendation service. The HNP is a Java-based tool that makes use of various Semantic Web technologies, and similar to most recommender systems, it operates based on user profiles. News messages are retrieved from RSS feeds, and processing steps make use of an OWL domain ontology that is constructed by domain experts. News items are classified using the GATE natural language processing software [8] and the WordNet [9] semantic lexicon.

4.1 Ceryx User Interface

The Ceryx plugin provides a tabbed graphical user interface, presenting the user an overview of all available news items from specified RSS feeds, the recommended news items based on a specified recommendation method, and evaluation results providing insights into the performance of the selected recommendation method. When browsing through the news items provided by the previously specified RSS feeds, the user is given an overview of news items, each of which is displayed with a title, date, and abstract. Furthermore, a Web browser is launched containing the full news message when an item is clicked, and the news item is added to the user profile. Users can choose from several recommendation methods, i.e., TF-IDF, SF-IDF, SF-IDF+, and various SS implementations. Resulting recommendation lists additionally display similarity scores ranging from 0 (low profile similarity) to 100 (best reflecting the user's interests), which are used for relevance sorting. Last, evaluation results are displayed for various performance measures. These results are calculated based on manually annotated test data.

4.2 Recommenders

We have implemented each of the mentioned recommendation methods in the Ceryx plugin. The traditional TF-IDF-based news recommendation implementation requires simple text processing: removal of stop words and reducing words to their lemma form.

The other, semantics-based, methods on the other hand require advanced text pre-processing. The original SF-IDF algorithm, as well as the SS variants operate using WordNet [9] synsets, which are obtained through part-of-speech tagging, stop word removal, lemmatisation, and word sense disambiguation. Part-of-speech tagging is performed by means of the Stanford Log-Linear Part-of-Speech Tagger [23]. Stop words (i.e., non-meaningful words) are removed by means of stop word list from the Onix Text Retrieval Toolkit API reference documents [18]. Next, lemmatisation is applied using the JAWS lemmatizer [22], in which the dictionary forms of words (lemmas) are determined that can be used for word lookup procedures in WordNet. Word sense disambiguation is performed using an implementation [14] of the adapted Lesk algorithm [1] and uses both the WordNet lemmas and identified parts-of-speech.

5. EVALUATION

In order to evaluate SF-IDF+ against its semantics-based alternatives and the TF-IDF baseline, we collected 100 news articles on technology companies from the Reuters RSS feed. Using information from three domain experts, we created a user profile based on item relatedness with respect to the eight given topics listed in Table 1: 'Asia or its countries', 'financial markets', 'Google & rivals', 'Web services', 'Microsoft & rivals', 'national economies', 'technology', and 'United States'. We employed a minimum inter-annotator agreement (IAA) of 66%, yet as depicted in Table 1, for each topic, the overall IAA was much higher. For each topic, the result set is split proportionally into a training set (60%) and a test set (40%). The user profile is created by adding all of the interesting news items from the training set.

Table 1: The number of interesting news items (*I*+), the number of non-interesting news items (*I*–), their associated inter-annotator agreements (*IAA*+ and *IAA*–, respectively), and the total inter-annotator agreement (*IAA*) for each topic.

Topic	I+	I–	IAA+	IAA–	IAA
Asia or its countries	21	79	100%	97%	99%
Financial markets	24	76	75%	68%	72%
Google & rivals	26	74	100%	95%	97%
Web services	26	74	96%	92%	94%
Microsoft & rivals	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%

5.1 Experimental Set-Up

In order to evaluate the SF-IDF+ recommendation method, we compare its performance to the performance of TF-IDF, SF-IDF, and SS in terms of F_1 scores, which are commonly used in this context, and hence are our main focus. Moreover, we also report on accuracy, precision, recall, and specificity. Performances are evaluated using an arbitrary cut-off value of 0.5 and using optimized cut-off values (based on F_1 scores, with the cut-off values ranging from 0.1 to 0.9 with an increment of 0.1). Additionally, we analyze graphs of F_1 scores over the full range of cut-off values and assess the significance of the results. We investigate whether a recommender performs significantly better than another recommender by employing a one-tailed two-sample paired Student t-test. With a specific level of significance, α, the null and alternative hypotheses are defined as:

$$H_0 : \mu_1 = \mu_2 \ , \ H_1 : \mu_1 > \mu_2 \ , \qquad (22)$$

where μ_1 is the mean performance on the F_1-measure of the first recommender and μ_2 is the mean performance on the F_1-measure of the second recommender.

Last, we optimize the weights used in SF-IDF+ using a genetic algorithm, which aims to maximize F_1-scores. The genetic algorithm is executed with a population of 333, a mutation probability of 0.1, elitism of 50, and a maximum number of 1,250 generations.

5.2 Experimental Results

After optimizing the SF-IDF+ weights with a cut-off of 0.5, we obtain the weights as shown in Table 2. The higher the weight of a specific semantic relation, the higher its importance. The semantic relationships with high weight values are the ‘member meronym’, ‘attribute’, and ‘domain of synset - region’. Given the investigated topics, one could expect a high weight for the ‘domain of synset - region’ relationship, as two out of eight queries specifically focus on physical areas. The other two important relationships are intuitive as well, because the related synsets with a ‘member meronym’ relationship are a part of the original synset in the news article, and the related synsets with the ‘attribute’ relationship are adjectives which express values of the original synset in the news article. Surprisingly, however, the importance of the semantic relationship ‘hypernym’ is rather low. This is most likely due to the fact that the synsets taken into account are too general and would merely act as a distraction to the recommender. Also the semantic relationship ‘hyponym’ does not have a high value. This result can be traced back to the fact that the specific synsets from the hyponym relationship do not appear often in the evaluated news articles, which is likely because of their specificity.

Table 2: Optimized semantic relationship weights based on F_1-score maximization for a cut-off of 0.5.

Relationship	Weight
Member meronym	0.981
Attribute	0.886
Domain of this synset - Region	0.828
Cause	0.747
Derivationally related form	0.739
Member of this domain - Topic	0.709
Domain of this synset - Usage	0.682
Member of this domain - Usage	0.667
Domain of this synset - Topic	0.570
Verb Group	0.518
Participle	0.493
Entailment	0.488
Substance holonym	0.475
Substance meronym	0.455
Antonym	0.416
Also see	0.380
Derived from adjective	0.374
Similar to	0.360
Part holonym	0.197
Pertainym	0.197
Hypernym	0.188
Member holonym	0.137
Instance hyponym	0.133
Instance hypernym	0.076
Part meronym	0.026
Member of this domain - Region	0.024
Hyponym	0.001

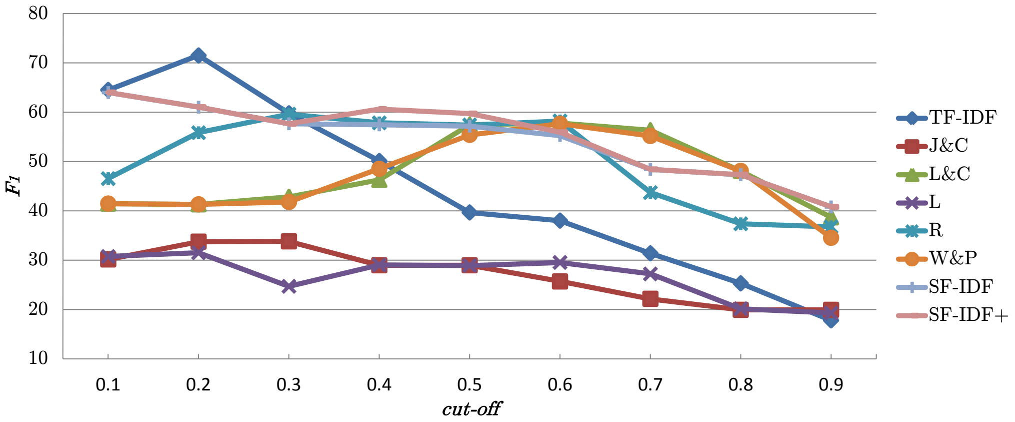

Figure 1: F_1-scores measured for the TF-IDF, Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), SF-IDF, and SF-IDF+ recommenders for various cut-off values, ranging from 0.1 to 0.9 with an increment of 0.1.

Table 3: Test results for TF-IDF, Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), SF-IDF, and SF-IDF+ in terms of accuracy, precision, recall, specificity, and F_1 for a cut-off of 0.5. Best values are printed in bold font.

	Accuracy	Precision	Recall	Specificity	F_1
TF-IDF	77.50%	**91.75%**	27.00%	**99.00%**	39.67%
J&C	67.38%	43.00%	23.88%	84.88%	28.96%
L&C	70.00%	63.38%	60.38%	73.63%	57.47%
L	61.13%	36.25%	26.25%	75.63%	28.90%
R	78.50%	76.75%	47.00%	91.75%	57.42%
W&P	66.13%	59.25%	**61.38%**	67.50%	55.42%
SF-IDF	79.13%	80.00%	45.38%	93.63%	57.14%
SF-IDF+	**79.75%**	80.13%	48.50%	93.13%	**59.73%**

After using these values as input for the SF-IDF+ based recommender, we obtained the results as presented in Table 3 for the TF-IDF based recommender, the various SS recommenders, the original SF-IDF recommender, and the SF-IDF+ recommender, while using a minimum cut-off value of 0.5. The Jiang & Conrath, Lin and TF-IDF recommenders perform rather poorly compared to the other recommenders. Based on our earlier work [6], one could expect that the original SF-IDF based recommender along with the Wu & Palmer would outperform both the Resnik and the Leacock & Chodorow recommenders. Surprisingly, however, taking into account other queries (compared to [6]), the opposite seem to be the case. Yet, with the adjusted SF-IDF, i.e., SF-IDF+, we outperform the original SF-IDF recommender by more than 2.5% with respect to the F_1-measure, which makes the SF-IDF+ based recommender the best performing recommendation method overall. In addition, we measure a better precision (80.13%) and recall (48.50%), and also the recommender's accuracy is increased by 0.65%, while giving up 0.5% on the recommender's specificity.

An overview of the p-values resulting from the one-tailed two-sample paired Student t-tests on F_1-scores is shown in Table 4. Depending on the employed confidence level, the p-value determining whether or not a news recommender significantly outperforms another recommender differs. The confidence levels typically used are 90%, 95%, and 99%, corresponding to p-values of 0.1, 0.05, and 0.01, respectively. While the SF-IDF+ news recommender outperforms all of the other news recommenders, it is not significantly better than the Leacock & Chodorow, Resnik, and the Wu & Palmer recommenders when using a confidence level of 90%. However, our SF-IDF+ news recommender performs significantly better than the original SF-IDF news recommender using a 90% confidence interval, and outperforms the TF-IDF-based news recommendation method even at a confidence interval of 95%. Last, our proposed semantics-based recommender is statistically better than the Lin and Jiang & Conrath news recommenders at a confidence level of 99%.

Using Figure 1 we determine the optimal cut-off value and corresponding optimal F_1-measure for every recommender. The most noteworthy difference with using a cut-off value of 0.5 is that the TF-IDF based news recommender outperforms both the original SF-IDF and the SF-IDF+ recommenders while using a relatively low cut-off value. However, this only holds for the low cut-off values between 0.1 and 0.3, since for every other minimum cut-off value the SF-IDF+ news recommendation method outperforms the TF-IDF-based method. The reason for this is that with a low minimum cut-off value the original SF-IDF recommender incorrectly discarded an interesting news article from the set of interesting news articles in six out of eight queries, which decreased its F_1 scores. After a thorough examination of the news articles involved, we found that in addition to the relative small size of the articles, there were also words that were not found in WordNet. This caused similarity values that were lower than expected, and, as a result, the TF-IDF news recommender performed better for that specific cut-off value than the SF-IDF and SF-IDF+ news recommenders.

Moreover, the SF-IDF+ news recommender does not outperform SF-IDF for low and high cut-off values. For low cut-off values, the news articles that have a similarity value higher than the cut-off values do not require the related synsets from the SF-IDF+ news recommender in order to be recommended. The higher cut-off values are simply too high for the SF-IDF+ news recommender to profit from the related synsets. Hence, the improvement of the SF-IDF+ news recommender over the SF-IDF news recommender is only visible at cut-off values between 0.4 and 0.6.

Table 4: One-tailed two-sample paired Student t-test p-values (bold when significant) for the F_1 values for the TF-IDF, Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), SF-IDF, and SF-IDF+ recommenders ($H_0 : \mu_{row} = \mu_{column}$, $H_1 : \mu_{row} > \mu_{column}$) for a cut-off of 0.5.

	TF-IDF	J&C	L&C	L	R	W&P	SF-IDF	SF-IDF+
TF-IDF	-	0.199	**0.934**	0.207	**0.938**	**0.924**	**0.938**	**0.950**
J&C	0.801	-	**0.989**	0.519	**0.985**	**0.977**	**0.984**	**0.992**
L&C	0.066	0.011	-	0.013	0.508	0.137	0.547	0.843
L	0.793	0.481	**0.987**	-	**0.984**	**0.975**	**0.983**	**0.991**
R	0.062	0.015	0.492	0.016	-	0.220	0.595	0.872
W&P	0.077	0.023	0.863	0.025	0.780	-	0.730	0.897
SF-IDF	0.062	0.016	0.453	0.017	0.405	0.270	-	**0.934**
SF-IDF+	0.050	0.008	0.157	0.009	0.128	0.103	0.066	-

Table 5: One-tailed two-sample paired Student t-test p-values (bold when significant) for the F_1 values for the TF-IDF, Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), SF-IDF, and SF-IDF+ recommenders ($H_0 : \mu_{row} = \mu_{column}$, $H_1 : \mu_{row} > \mu_{column}$) for a cut-off optimized per recommender.

	TF-IDF	J&C	L&C	L	R	W&P	SF-IDF	SF-IDF+
TF-IDF	-	0.002	0.015	0.000	0.042	0.016	0.115	0.115
J&C	**0.998**	-	**0.962**	0.203	**0.966**	**0.966**	**0.992**	**0.992**
L&C	**0.985**	0.038	-	0.014	0.838	0.453	**0.925**	**0.925**
L	**1.000**	0.797	**0.986**	-	**0.986**	**0.988**	**0.998**	**0.998**
R	**0.958**	0.034	0.162	0.014	-	0.162	0.808	0.808
W&P	**0.984**	0.034	0.457	0.012	0.838	-	**0.929**	**0.929**
SF-IDF	0.885	0.008	0.075	0.002	0.192	0.071	-	-
SF-IDF+	0.885	0.008	0.075	0.002	0.192	0.071	-	-

When we evaluate the performance of the news recommendation methods when selecting the optimal cut-off value for each method individually, we observe that despite the fact that the TF-IDF news recommender outperforms the SF-IDF and SF-IDF+ news recommendation methods in terms of F_1, the differences are not statistically significant, even when using a 90% confidence interval as shown in Table 5. Another difference between using the cut-off value of 0.5 and the individual optimal cut-off values is that the SF-IDF+ news recommender is now significantly better than the Leacock & Chodorow and Wu & Palmer news recommenders. This is in contrast to the Resnik recommender, which now obtains a better p-value with respect to our recommender. Because all F_1-scores that make up the optimal F_1-scores for the SF-IDF news recommender and the SF-IDF+ news recommender are the same, the corresponding p-value could not be computed.

The overall average performance of every news recommender depicted in Figure 1 shows the benefits of employing SF-IDF+ recommendation over other (semantics-based) recommendation methods. When taking into account the full range of the evaluated cut-off values, we obtain p-values as shown in Table 6. From this table, we derive that the Jiang & Conrath, Leacock & Chodorow, and the Lin news recommenders are all significantly outperformed by our SF-IDF+ news recommender at a confidence level of 99%. The TF-IDF-based recommender and the Resnik and Wu & Palmer recommenders on the other hand are significantly outperformed by our SF-IDF+ recommender at a confidence level of 95%, while the original SF-IDF news recommender is significantly outperformed at a confidence level of 90%.

6. CONCLUSION

In most content-based news recommendation platforms, recommendation is performed using the TF-IDF weighting scheme combined with a cosine similarity measure. In order to better cope with news information, we have looked into semantics-driven methods, that take into account term meaning by exploiting semantic lexicon synsets and the cosine similarity (SF-IDF) or by using semantic similarities (SS). However, such systems do not take into account the various inter-synset semantic relationships providing only for a limited understanding of news semantics.

We explored the feasibility of extending the SF-IDF news recommendation method, in order to additionally account for the semantic relations between synsets in a semantic lexicon. The proposed method, SF-IDF+, has been implemented in Ceryx, an extension to the Hermes news personalization service. Our evaluation on 100 financial news messages and 8 topics showed that, on average, SF-IDF+ outperforms the other methods in terms of F_1-scores (as discretized in the paper), and works best for the most commonly used cut-off values around 0.5, associated with systems having a high precision without compromising too much on recall. Moreover, SF-IDF+ is more stable (i.e., less dependent on a specific cut-off value) than the other methods.

The discussed recommenders are based on semantic lexicon synsets. However, such recommenders are dependent on the information available in such lexicons. Hence, as future work, we would like to explore a way to combine multiple lexicons. Moreover, we would like to create an expert system for collecting and updating information. Also, we would like to investigate other learning techniques for optimizing semantic relationship weights. Last, one could research a

Table 6: One-tailed two-sample paired Student t-test p-values (bold when significant) for the average F_1 values for the TF-IDF, Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), SF-IDF, and SF-IDF+ recommenders ($H_0 : \mu_{row} = \mu_{column}$, $H_1 : \mu_{row} > \mu_{column}$) for all cut-offs.

	TF-IDF	J&C	L&C	L	R	W&P	SF-IDF	SF-IDF+
TF-IDF	-	0.003	0.680	0.005	0.881	0.651	**0.982**	**0.985**
J&C	**0.997**	-	**1.000**	0.425	**1.000**	**1.000**	**1.000**	**1.000**
L&C	0.320	0.000	-	0.000	0.758	0.132	**0.995**	**0.996**
L	**0.995**	0.575	**1.000**	-	**1.000**	**1.000**	**1.000**	**1.000**
R	0.119	0.000	0.242	0.000	-	0.179	**0.948**	**0.976**
W&P	0.349	0.000	0.868	0.000	0.821	-	**0.966**	**0.976**
SF-IDF	0.018	0.000	0.005	0.000	0.052	0.034	-	**0.937**
SF-IDF+	0.015	0.000	0.004	0.000	0.024	0.024	0.063	-

means to cope with named entities that are not available in a lexicon, e.g., through Web search page co-counts.

Acknowledgment

The authors are partially supported by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and the Dutch national program COMMIT.

7. REFERENCES

[1] 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)

[2] Bharat, K., Kamba, T., Albers, M.: Personalized, Interactive News on the Web. Multimedia Systems 6(5), 349–358 (1998)

[3] Billsus, D., Pazzani, M.J.: A Personal News Agent that Talks, Learns and Explains. In: Etzioni, O., Müller, J.P., Bradshaw, J.M. (eds.) 3rd Annual Conference on Autonomous Agents (AGENTS 1999). pp. 268–275. ACM (1999)

[4] 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)

[5] Cantador, I., Bellogín, A., Castells, P.: Ontology-Based Personalised and Context-Aware Recommendations of News Items. In: 2008 IEEE / WIC / ACM International Conference on Web Intelligence (WI 2008). pp. 562–565. IEEE Computer Society (2008)

[6] 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)

[7] Carreira, R., Crato, J.M., Gonçalves, D., Jorge, J.A.: Evaluating Adaptive User Profiles for News Classification. In: 9th International Conference on Intelligent User Interfaces (IUI 2004). pp. 206–212. ACM (2004)

[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] Frasincar, F., Borsje, J., Levering, L.: A Semantic Web-Based Approach for Building Personalized News Services. International Journal of E-Business Research 5(3), 35–53 (2009)

[11] Frasincar, F., IJntema, W., Goossen, F., Hogenboom, F.: Business Intelligence Applications and the Web: Models, Systems and Technologies, chap. A Semantic Approach for News Recommendation, pp. 102–121. IGI Global (2011)

[12] Getahun, F., Tekli, J., Chbeir, R., Viviani, M., Yetongnon, K.: Relating RSS News/Items. In: Gaedke, M., Grossniklaus, M., Díaz, O. (eds.) 9th International Conference on Web Engineering (ICWE 2009). pp. 442–452. Springer-Verlag (2009)

[13] 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 Thirteenth International Conference on Database Theory (EDBT/ICDT 2010). ACM (2010)

[14] 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)

[15] Jiang, J.J., Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: 10th International Conference on Research in Computational Linguistics (ROCLING 1997). pp. 19–33 (1997)

[16] Lang, K.: NewsWeeder: Learning to Filter Netnews. In: 12th International Conference on Machine Learning (ICML 1995). pp. 331–339. Morgan Kaufmann (1995)

[17] Leacock, C., Chodorow, M.: WordNet: An Electronic Lexical Database, chap. Combining Local Context and WordNet Similarity for Word Sense Identification, pp. 265–283. MIT Press (1998)

[18] Lextek: Onix Text Retrieval Toolkit – API Reference. `http://www.lextek.com/manuals/onix/stopwords1.html` (2012)

[19] Lin, D.: An Information-Theoretic Definition of Similarity. In: Shavlik, J.W. (ed.) 15th International Conference on Machine Learning (ICML 1998). pp. 296–304. Morgan Kaufmann (1998)

[20] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: 14th International Joint Conference on Artificial Intelligence (IJCAI 1995). pp. 448–453. Morgan Kaufmann (1995)

[21] Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)

[22] Spell, B.: Java API for WordNet Searching (JAWS). `http://lyle.smu.edu/~tspell/jaws/index.html` (2012)

[23] 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)

[24] Wu, Z., Palmer, M.S.: Verb Semantics and Lexical Selection. In: 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994). pp. 133–138. Association for Computational Linguistics (1994)