Bing-SF-IDF+: A Hybrid Semantics-Driven News Recommender

Michel Capelle
michelcapelle@gmail.com

Marnix Moerland
marnix.moerland@gmail.com

Frederik Hogenboom
fhogenboom@ese.eur.nl

Flavius Frasincar
frasincar@ese.eur.nl

Damir Vandic
vandic@ese.eur.nl

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

ABSTRACT

Content-based news recommendation is traditionally performed using the cosine similarity and the TF-IDF weighting scheme for terms occurring in news messages and user profiles. Semantics-driven variants such as SF-IDF additionally take into account term meaning by exploiting synsets from semantic lexicons. However, they ignore the various semantic relationships between synsets, providing only for a limited understanding of news semantics. Moreover, semantics-based weighting techniques are not able to handle – often crucial – named entities, which are often not present in semantic lexicons. Hence, we extend SF-IDF by also considering the synset semantic relationships, and by employing named entity similarities using Bing page counts. Our proposed method, Bing-SF-IDF+, outperforms TF-IDF and SF-IDF in terms of F_1-scores and kappa statistics based on a news data set.

1. INTRODUCTION

Nowadays, the Web has become an important source of information, housing more information providers than ever before. One of the Web's main contents is media, mostly in the form of news. However, the user is confronted with an overload of information, and hence, many recommendation methods have been developed to filter and structure this information. Recommender systems lend a hand in distinguishing between interesting and non-interesting products, news articles, etc. Based on user preferences or characteristics, captured in elicited or derived user profiles, recommendations can be made. There are three basic types of recommendation systems: content-based recommenders, which recommend news items based on their content, collaborative filtering recommenders, which recommend news items by means of user similarity, and hybrid recommenders, that combine the previous two approaches.

Traditionally, content-based recommender systems are term-based, and hence operate on term frequencies. A commonly used measure is Term Frequency – Inverse Document Frequency (TF-IDF). When employing user profiles that describe users' interest based on previously browsed items, these can be translated into vectors of TF-IDF weights. With a measure like cosine similarity, one can calculate the interestingness of a new item (with respect to a user profile). For this, TF-IDF weights are computed on every term within a document.

However, TF-IDF-based systems and the like do not consider the text semantics, which could be added by using Web ontologies, yet these are mostly domain dependent, requiring continuous maintenance. One could also employ synonym sets (synsets) from general semantic lexicons (e.g., WordNet [9]), eliminating the need for domain ontologies. A synset is a collection of one or more words sharing parts-of-speech and meaning. For example, the synset 'turkey' and 'meleagris gallopavo' refers to an animal, while the description of the synset consisting of the words 'Turkey' and 'Republic of Turkey' describes a country. While the words are identical, their synsets and thus their meanings are not the same. Using synsets as proxies for document semantics could hence prove to be a viable addition to recommendation methods.

In previous work [5], we hence introduced the Synset Frequency – Inverse Document Frequency (SF-IDF) measure, operating on WordNet synsets instead of terms. We evaluated SF-IDF with respect to TF-IDF and a semantics-based alternative, Semantic Similarity (SS), and showed the benefits of considering synsets. However, up until now, we did not take into account inter-synset relationships. For example, the 'turkey, Meleagris gallopavo' synset is a hyponym of the synset consisting of the words 'domestic fowl', 'fowl', and 'poultry' referring to a bird. 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 [11].

Additionally, a vast amount of named entities appear in news articles, which could provide crucial information when constructing user profiles. When employing a synset-based

SAC'15 April 13 - 17 2015, Salamanca, Spain
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3196-8/15/04 ...$15.00
http://dx.doi.org/10.1145/2695664.2695700

method for (news) recommendation, named entities are usually not taken into consideration. One could enhance existing semantics-based recommendation methods by employing similarities based on page counts gathered by Web search engines, e.g., Google or Bing.

Therefore, we extend SF-IDF by additionally considering synset semantic relationships, and by using named entity similarities using Bing page counts. The proposed recommendation method, Bing-SF-IDF+, as well as SF-IDF and several semantic lexicon-driven similarity methods are implemented and evaluated. The remainder of this paper is organized as follows. First, we discuss related work in Section 2. Next, we introduce the hybrid semantics-driven Bing-SF-IDF+ news recommender in Section 3. Last, we evaluate our performance in Section 4 and draw conclusions in Section 5.

2. RELATED WORK

Many profile-based recommender systems exist, differing in their approaches for news recommendation. Most importantly, they implement different similarity measures for calculating similarities between a news item and a user profile, i.e., the set of news items previously read by a user.

2.1 TF-IDF

One of the most commonly used similarity measures is TF-IDF, combined with cosine similarities. The TF-IDF method is composed of two parts: term frequency $tf(t, d)$ and inverse document frequency $idf(t, d)$, and operates on terms T in documents D. The term frequency 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 expresses the inverse of the occurrence of a term t in a set of documents D and is defined as:

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

where $|D|$ is the total amount of documents in the set of documents being compared to one another, and $d : t \in d$ denotes the amount of documents which contain term t. When multiplying $tf(t, d)$ and $idf(t, d)$, we obtain $tf\text{-}idf(t, d)$:

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

For every term t in document d, the TF-IDF value is computed and stored in a vector $A(d)$. This computation is performed for all documents in D. Subsequently, the similarity between a set of terms from news item d_u and a user profile d_r is calculated using the cosine similarity measure:

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

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 the cut-off value are recommended to the user.

2.2 SF-IDF

A drawback of TF-IDF is that word semantics are not taken into account, resulting in different words with the same meaning to be counted as two separate terms, and a word appearing for two different meanings to be counted as one. Therefore, semantics-based similarity measures have been proposed. A TF-IDF variant is the Synset Frequency – Inverse Document Frequency (SF-IDF) [5], which makes use of synonym sets (synsets) from a semantic lexicon (e.g., WordNet [9]) instead of terms. These synsets are obtained after performing word sense disambiguation. After replacing term t by synset s, the SF-IDF formulas are:

$$sf\text{-}idf(s, d) = sf(s, d) \times idf(s, d) \ , \tag{5}$$

$$sim_{sf\text{-}idf}(d_u, d_r) = \frac{A(d_u) \cdot A(d_r)}{||A(d_u)|| \times ||A(d_r)||} \ . \tag{6}$$

Similar to the TF-IDF method, the cosine similarity measure is used to compute the similarity of an unread news article to the user profile, and the unread news items with a rating above the cut-off are suggested to the user.

2.3 Semantic Similarity

Another semantics-based measure is Semantic Similarity (SS) [5], in which one compares synsets from the unread news items with synsets from all the news items in the user profile. This is performed by employing pairs between the elements of the two sets with a common part-of-speech. In order to do so, we define V, which contains all the possible pairs of synsets from the unread news item d_u, U, and the union of synsets from the user profile d_r, R:

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

where u_k represents a synset from the unread news item, r_l represents a synset from the user profile, k is the number of synsets in the unread news item, and l is the number of synsets in the user profile. A subset of V that contains all the combinations that have a common part-of-speech is described as:

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

where $POS(u)$ 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 (gathered from a semantic lexicon as WordNet). The final similarity rank is defined as:

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

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

Some similarity measures are based on the information content of the corresponding nodes. The information content (IC) of a node is the negative logarithm of the sum of all probabilities of all the words in the synset:

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

where $p(x)$ denotes the probability that an instance x of synset s occurs in a corpus, and w represents a word sense in synset s. The Jiang & Conrath [14] measure uses the information content of both the synsets and the lowest common subsumer (LCS), while Lin's measure [16] makes use of the logarithms of the chances of appearance of both and the lowest common subsumer. Resnik's measure [17] on the other hand uses the information content of the lowest common subsumer of the two synsets. The information content-based measures are defined as follows:

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

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

$$sim_{R}(u,r) = IC(LCS(u,r)) \ . \tag{13}$$

The Leacock & Chodorow [15] and Wu & Palmer [19] measures, on the other hand, make use of the path length between the nodes. 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 is based on the shortest path length Λ between nodes u and r (where Ω is the maximum depth of the taxonomy). The Wu & Palmer's similarity measure on the other hand makes use of the depth of the lowest common subsumer of both nodes and the path length between them. The measures are defined as follows:

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

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

2.4 Improvements

Even though the SF-IDF and SS methods have a notion of semantics, they do not take into account semantic relations. However, research has shown that the relationships between concepts provide more structure in a text and hence contribute to an improved level of interpretability [11]. The authors of [11] propose using Semantic Relatedness to recommend news articles to the user by making use of WordNet relationships regarding synonymy, hyponymy, and meronomy. Weights are assigned based on maximum enclosure similarities. In our work, we also make use of WordNet relationships, yet we do not limit ourselves to a subset of relations, but instead utilize all relationships available in WordNet. Furthermore, we aim for a more advanced mechanism for determining importance weights for these relationships in the form of a machine learning approach.

Complementary, one could enhance existing semantics-based recommendation methods by employing similarities based on page counts gathered by Web search engines for named entities. Typically, these named entities do not appear in a semantic lexicon and thus cannot be covered by the previous approach. The more a pair of entities co-occur on Web sites, the more likely it is that there is a similarity between both entities [3]. A frequently studied similarity measure based on page counts is the Normalized Google Distance (NGD) [6], which is a normalized semantic distance with values between 0 and 1 that is calculated using probabilities related to the number of hits associated with the two separate entities, the number of hits associated with the two entities appearing together, and the number of indexed Web pages. However, as at the time of writing, Google's API was not available as a free service anymore, in our research we made use of Bing, which still offered an API for free [2].

3. BING-SF-IDF+ RECOMMENDATION

As is the case for most semantics-based news recommendation methods, Bing-SF-IDF+ performs recommendations based on a user profile, reflecting the user's interests. Under the assumption that users only read news items to their likings, a user profile consists of all currently read news items. A user profile is updated upon reading previously unseen news items by the associated user. In order to perform news recommendation, for every unread news item, a similarity score between the news article and the user profile is computed. Unread news items having a similarity score that exceed a predefined cut-off value are recommended to the user.

The Bing-SF-IDF+ similarity score is a weighted average of two separate similarity scores, each expressing a different type of similarity. The Bing component expresses similarities between named entities, whereas SF-IDF+ measures the similarities between synsets. This section continues by describing the calculation of both similarity scores, and subsequently we explain the combination of both scores into the final similarity score.

3.1 Bing

The Bing similarity score takes into account the named entities which do not occur in a semantic lexicon (e.g., WordNet [9]). These named entities are derived from news articles through a named entity recognizer (e.g., from the Alias-i LingPipe [1] software). We describe an unread news item d_u and the user profile d_r using sets of named entities, U and R, respectively:

$$U = \{u_1, u_2, \ldots, u_k\} \ , \tag{16}$$

$$R = \{r_1, r_2, \ldots, r_l\} \ , \tag{17}$$

where u_k represents a named entity in the unread news item U, r_l denotes a named entity in the user profile R, and k and l are the number of named entities in the unread news item and in the user profile, respectively.

Next, we construct a vector containing all possible pairs of named entities from the unread news item d_u and the user profile d_r:

$$V = (\langle u_1, r_1 \rangle, \ldots, \langle u_k, r_l \rangle) \ \forall \ u \in U, \ r \in R \ . \tag{18}$$

Subsequently, we use search engine page counts of the named entity pairs in order to measure the similarity between the pairs. The page count is defined as the number of Web pages that were found by the Bing Web search engine that contain a named entity or a pair of named entities. For every pair (u, r) in V we compute the page rank-based Point-Wise Mutual Information (PMI) co-occurrence similarity measure [4] instead of the NGD [6], due to the unavailability of the Google API. PMI, in our case, is a measure of association between two probabilities, measuring the difference between the actual and expected joint probability of the occurrence of two named entities in a query on a Web search engine, based on the marginal probabilities of the two named entities while assuming independence. The PMI similarity measure sim_{PMI} for pair (u, r) is defined as:

$$sim_{PMI}(u, r) = \log \frac{\frac{c(u,r)}{N}}{\frac{c(u)}{N} \times \frac{c(r)}{N}} \ , \qquad (19)$$

where $c(u, r)$ is the page count for the pair (u, r) of named entities, $c(u)$ and $c(r)$ are the page counts for the named entities u from the unread news item and r from the user profile, respectively, and N denotes the number of Web pages that are indexed by the Bing Web search engine, which is approximately 15 billion [12].

Last, the Bing similarity score is defined as the average of the PMI similarity scores over all named entity pairs:

$$sim_{Bing}(V) = \frac{\sum\limits_{(u,r)\in V} sim_{PMI}(u, r)}{|V|} \ . \qquad (20)$$

3.2 SF-IDF+

The SF-IDF+ similarity score takes into account sets of synonyms (synsets) of words that occur in a semantic lexicon (e.g., WordNet [9]) and is based on the Synset Frequency – Inverse Document Frequency similarity measure introduced in earlier work [5]. First, all synsets are retrieved from the unread news items by employing natural language processing techniques. The set of synsets is extended by appending the concepts that are referred to by semantical relationships of the included synsets, and hence is then defined as:

$$S(s) = \{s\} \cup \bigcup_{r \in R(s)} r(s) \ , \qquad (21)$$

where s is the synset in the news item, $r(s)$ is the synset that is related to synset s by relationship r, and $R(s)$ is the set of relationships of synset s.

The unread news item and the user profile can be described as sets of extended synsets:

$$U = \{S(u_1), S(u_2), \ldots, S(u_k)\} \ , \qquad (22)$$

$$R = \{S(r_1), S(r_2), \ldots, S(r_l)\} \ , \qquad (23)$$

where $S(u_k)$ 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, yet SF-IDF+ makes use of extended synsets instead of terms (TF-IDF) or synsets (SF-IDF), and weighting is applied depending on the relationship that the semantically related synset has with the synset. We define the SF-IDF+ weight for the unread news item d_u and the user profile d_r as:

$$sf\text{-}idf{+}(s, d, r) = sf(s, d) \times idf(s, d) \times w_r \ , \qquad (24)$$

with $d \in \{d_u, d_r\}$, where $sf(s, d)$ is 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, and w_r is the weight of the relationship r between the semantically related synset and the synset s, which can be optimized, for example, by means of a genetic algorithm.

Then, two vectors are constructed, representing the unread news item d_u and the user profile d_r, each containing all $sf\text{-}idf(s, d)$ and $sf\text{-}idf{+}(s, d, r)$ weights for all extended synsets s in d:

$$A(d) = \begin{Bmatrix} \varsigma(s_1, d), \varsigma(s_1, d, r_1), \ldots, \varsigma(s_1, d, r_{m_{s_1}}), \\ \varsigma(s_2, d), \varsigma(s_2, d, r_1), \ldots, \varsigma(s_2, d, r_{m_{s_2}}), \\ \cdots \\ \varsigma(s_n, d), \varsigma(s_n, d, r_1), \ldots, \varsigma(s_n, d, r_{m_{s_n}}) \end{Bmatrix} \ , \qquad (25)$$

where $\varsigma(s, d)$ represents $sf\text{-}idf(s, d)$, $\varsigma(s, d, r)$ represents $sf\text{-}idf{+}(s, d, r)$, n denotes the total number of synsets in document d, and m_{s_i} is the total number of synsets related to synset s_i.

Last, we compute the similarity score between the unread news item d_u and the user profile d_r with the cosine similarity measure that is defined as:

$$sim_{sf\text{-}idf+}(d_u, d_r) = \frac{A(d_u) \cdot A(d_r)}{||A(d_u)|| \times ||A(d_r)||} \ . \qquad (26)$$

3.3 Bing-SF-IDF+

For every unread news item d_u, we now have a Bing similarity score sim_{Bing} and a SF-IDF+ similarity score $sim_{sf\text{-}idf+}$. We normalize the similarity scores of both components in order to make both scores compatible. For this, we employ min-max normalization between 0 and 1 on both sets of similarity scores:

$$\overline{sim}_{Bing}(d_u, d_r) = \frac{sim_{Bing}(d_u, d_r) - \min\limits_{u} sim_{Bing}(d_u, d_r)}{\max\limits_{u} sim_{Bing}(d_u, d_r) - \min\limits_{u} sim_{Bing}(d_u, d_r)} \ , \qquad (27)$$

$$\overline{sim}_{sf\text{-}idf+}(d_u, d_r) = \frac{sim_{sf\text{-}idf+}(d_u, d_r) - \min\limits_{u} sim_{sf\text{-}idf+}(d_u, d_r)}{\max\limits_{u} sim_{sf\text{-}idf+}(d_u, d_r) - \min\limits_{u} sim_{sf\text{-}idf+}(d_u, d_r)} \ . \qquad (28)$$

The final Bing-SF-IDF+ similarity score is computed by taking a weighted average of the normalized similarity scores of the Bing and SF-IDF+ elements:

$$sim_{Bing\text{-}sf\text{-}idf+}(d_u, d_r) = \alpha \times \overline{sim}_{Bing}(d_u, d_r) + (1 - \alpha) \times \overline{sim}_{sf\text{-}idf+}(d_u, d_r) \ , \qquad (29)$$

where α is a weight that is optimized on a training set. All the unread news items which have a similarity score that exceeds the predefined cut-off value are recommended to the user.

3.4 Bing-SF-IDF+ Implementation

Our framework is implemented as an extension to the Ceryx [5] plugin of the Hermes News Portal (HNP) [10], a news recommendation service. The HNP, a Java-based tool that makes use of various Semantic Web technologies, operates based on user profiles and processes news items from RSS feeds. The core of the HNP is an OWL domain ontology that is constructed by domain experts, allowing for semantics-based operations on news messages. For this paper we did not make use of the OWL domain ontology of Hermes. These items are classified using the GATE natural language processing software [8] and the WordNet [9]

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 and its rivals	26	74	100%	95%	97%
Web services	26	74	96%	92%	94%
Microsoft and its 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%

semantic lexicon. The semantics-based methods additionally make use of the Stanford Log-Linear Part-of-Speech Tagger [18], Lesk Word Sense Disambiguation [13], and the Alias-i's LingPipe 4.1.0 [1] Named Entity Recognizer. Page counts are gathered through the Bing API 2.0 [2].

4. EVALUATION

In order to evaluate Bing-SF-IDF+ against its semantics-based alternatives and TF-IDF, we collected a data set containing 100 news articles from a Reuters news feed on technology companies. Three users from our university with expertise in news analytics indicated whether a news article relates to one of eight given topics. Out of these user ratings, a user profile was constructed for every topic using a minimum inter-annotator agreement (IAA) of 66%. Table 1 displays the resulting number of interesting and non-interesting news items per topic, as well as their associated agreements and the total inter-annotator agreement. For each topic, the result set is split proportionally into a training set (60%) for creating the user profile and a test set (40%) for evaluation.

4.1 Experimental Set-Up

In order to evaluate the Bing-SF-IDF+ recommendation method, we compare its performance to the performance of TF-IDF recommendation, the original SF-IDF method, and the five SS recommendation methods introduced earlier in terms of F_1 and kappa statistics [7] (measuring whether the proposed classification is better than a random guess), 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 for the individual topics using various cut-off values (i.e., items with similarity scores above a specific value are recommended), ranging from 0 to 1 with an increment of 0.01. Additionally, we analyze graphs of F_1 and kappa statistics over the full range of cut-off values and assess the significance of the results using a one-tailed two-sample paired Student t-test with a level of 95% significance. Last, we optimize the weights and the α-value used in Bing-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. These settings have been determined during initial experiments on a small, yet representative portion of our training set.

Experiments are run on the Lisa system, a SARA Computing and Networking Services cluster computer consisting of several hundreds of multi-core nodes running the Debian Linux AMD64 operating system. The computers employed for our experiments each have dodeca-core CPUs with 12MB cache running at 2.26GHz, and operate on 24GB of QPI 5.86 GT/s memory.

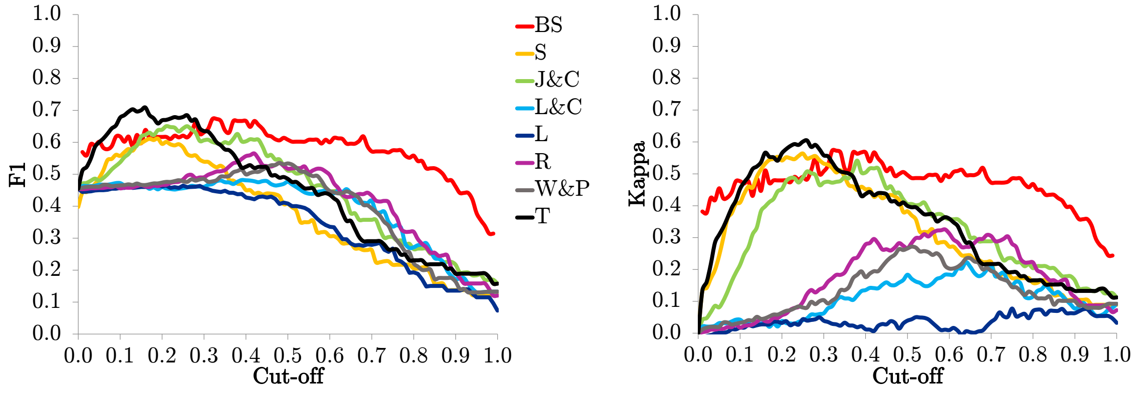
4.2 Experimental Results

Table 2 displays the average performance for each of the evaluated recommender methods (rows) in terms of accuracy, precision, recall, F_1, specificity, and kappa statistics (columns). Bing-SF-IDF+ outperforms TF-IDF and all semantic recommenders. The Jiang & Conrath recommender also shows good overall performance. The graphs in Figures 1(a) and 1(b), which provide a closer look into the F_1-scores and kappa statistics, support these findings. Figure 1(a) shows that for high cut-off values (i.e., above 0.3), Bing-SF-IDF+ outperforms all other recommenders in terms of F_1, while there is not much difference in performance between the other recommenders. For low cut-off values, TF-IDF performs best, and the Jiang & Conrath recommender performs good as well over the full range of values.

According to its definition, the Kappa statistic plotted in Figure 1(b) measures whether the proposed classifications are better than random guessing. The closer to 1, the more classification power a recommender has. Negative values indicate that a recommender performed worse than the expected performance with random guessing. Figure 1(b) demonstrates that for high cut-off values, the Bing-SF-IDF+ recommender scores a higher Kappa statistic than the TF-IDF recommender and the other semantic recommenders. This is an indication that the Bing-SF-IDF+ recommender seems to have more classification power than the other recommenders. Not only Bing-SF-IDF+ and TF-IDF show a

Table 2: Average test results for Bing-SF-IDF+ (BS), SF-IDF (S), Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), and TF-IDF (T).

	Acc.	Prec.	Rec.	F_1	Spec.	Kappa
BS	0.81	0.71	0.53	0.58	0.91	0.47
S	0.65	0.68	0.43	0.37	0.76	0.32
J&C	0.72	0.73	0.48	0.45	0.82	0.31
L&C	0.55	0.44	0.58	0.39	0.54	0.11
L	0.51	0.38	0.53	0.34	0.51	0.03
R	0.60	0.55	0.57	0.42	0.61	0.17
W&P	0.57	0.46	0.59	0.40	0.55	0.13
T	0.75	0.83	0.44	0.45	0.88	0.34



(a) F_1-scores.

(b) Kappa statistics.

Figure 1: F_1-scores and kappa statistics measured for the Bing-SF-IDF+ (BS), SF-IDF (S), Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), and TF-IDF (T) recommenders for various cut-off values.

good performance, but also the SF-IDF and Jiang & Conrath SS methods.

An overview of the p-values resulting from the one-tailed two-sample paired Student t-tests on F_1-scores and kappa statistics is shown in Tables 3 and 4. With a level of 95% significance, Bing-SF-IDF+ significantly outperforms all other approaches. Also, the tables support the conclusions that TF-IDF significantly outperforms SF-IDF, Leacock & Chodorow, Lin, Resnik, and Wu & Palmer, yet it does not significantly outperform Jiang & Conrath in terms of F_1. The difference between the performance of Bing-SF-IDF+ and SF-IDF is evident, as SF-IDF significantly outperforms only one other recommender, i.e., Lin SS. The Jiang & Conrath recommender performs a lot better, and outperforms all recommenders but Bing-SF-IDF+ and TF-IDF significantly.

For the kappa statistics, the results are more clear cut. Bing-SF-IDF+ significantly outperforms all other recommenders, and TF-IDF outperforms all but Bing-SF-IDF+ recommenders. SF-IDF also outperforms many of the evaluated recommenders, i.e., Leacock & Chodorow, Lin, Resnik, and Wu & Palmer. The other recommenders perform notably worse, with the Lin SS recommender being the worst performing recommender, significantly outperformed by all other recommendation methods.

Table 3: One-tailed two-sample paired Student t-test p-values for the F_1-measure averages for the Bing-SF-IDF+ (BS), SF-IDF (S), Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), and TF-IDF (T) recommenders ($H_0 : \mu_{column} = \mu_{row}$, $H_1 : \mu_{column} > \mu_{row}$, $\alpha = 0.05$).

	BS	S	J&C	L&C	L	R	W&P	T
BS		1.00	1.00	1.00	1.00	1.00	1.00	1.00
S	0.00		0.00	0.03	1.00	0.00	0.01	0.00
J&C	0.00	1.00		1.00	1.00	1.00	1.00	0.59
L&C	0.00	0.97	0.00		1.00	0.00	0.07	0.00
L	0.00	0.00	0.00	0.00		0.00	0.00	0.00
R	0.00	1.00	0.00	1.00	1.00		1.00	0.01
W&P	0.00	0.99	0.00	0.93	1.00	0.00		0.00
T	0.00	1.00	0.41	1.00	1.00	0.99	1.00	

Table 4: One-tailed two-sample paired Student t-test p-values for the kappa statistic averages for the Bing-SF-IDF+ (BS), SF-IDF (S), Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), and TF-IDF (T) recommenders ($H_0 : \mu_{column} = \mu_{row}$, $H_1 : \mu_{column} > \mu_{row}$, $\alpha = 0.05$).

	BS	S	J&C	L&C	L	R	W&P	T
BS		1.00	1.00	1.00	1.00	1.00	1.00	1.00
S	0.00		0.56	1.00	1.00	1.00	1.00	0.00
J&C	0.00	0.44		1.00	1.00	1.00	1.00	0.00
L&C	0.00	0.00	0.00		1.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00		0.00	0.00	0.00
R	0.00	0.00	0.00	1.00	1.00		1.00	0.00
W&P	0.00	0.00	0.00	1.00	1.00	0.00		0.00
T	0.00	1.00	1.00	1.00	1.00	1.00	1.00	

Table 5: Typical example of optimized semantic relationship weights based on F_1-score maximization for a cut-off of 0.33, using an optimized α of 0.35.

Relationship	Weight
Derived from adjective	1.00
Attribute	0.96
Instance hyponym	0.83
Substance holonym	0.82
Member meronym	0.74
Member of this domain - Usage	0.74
Derivationally related form	0.73
Part meronym	0.58
Domain of this synset - Topic	0.42
Participle	0.41
Member of this domain - Region	0.35
Pertainym	0.33
Member of this domain - Topic	0.31
Member holonym	0.28
Domain of this synset - Usage	0.24
Instance hypernym	0.22
Substance meronym	0.21
Similar to	0.20
Domain of this synset - Region	0.13
Cause	0.12
Also see	0.09
Part holonym	0.07
Verb group	0.06
Antonym	0.05
Entailment	0.03
Hypernym	0.01
Hyponym	0.01

Last, an evaluation of the optimized weights and α-values for all cut-off values leads to various insights. For Bing-SF-IDF+, scores are weighted using an average optimized α of 0.48 (with a standard deviation of 0.27), giving a substantial weight to Bing similarities as well as to the extended synsets incorporating semantic relationships, underlining the importance of both proposed extensions. Table 5 presents a typical example of the weights for each synset relationship, and uses an α of 0.35. The relationships that typically obtain high weights are 'attribute', 'derivationally related form', 'derived from adjective', 'instance hyponym', 'substance holonym', 'member meronym', and 'member of this domain - usage'.

Related synsets with a 'member meronym' or 'member of this domain - usage' relationship are a part of and thus are strongly semantically related to the original synset in the news article. Related synsets with the 'attribute' relationship are adjectives which often express members of the original synset in the news article. Also, 'derivationally related form' and 'derived from adjective' show strong connections between synsets and occur frequently in the employed semantic lexicon. Furthermore, 'instance hyponym' relations (stating a synset is an instance of another synset) have high weights, as these would typically lead to closely related, more abstract synsets, the utilization of which would improve user profile and news message matching. Last, the 'substance holonym' relation (expressing a synset is a substance and part of another synset) is a surprising relationship that obtained high weights, possibly due to the topics used in the evaluation.

5. CONCLUSIONS

In most recommendation applications, news recommendation is performed using the cosine similarity and the TF-IDF weighting scheme. In order to better cope with news information, recently, semantics-driven methods have been developed, taking into account term meaning by exploiting semantic lexicon synsets and the cosine similarity (SF-IDF) or by making use of semantic (lexicon-driven) similarities (SS). However, such systems do not take into account the various semantic relationships between synsets, like synonymy, homonymy, etc., providing only for a limited understanding of news semantics. Additionally, named entities are not considered, as these are often not present in semantic lexicons.

In this paper, we explored the possibilities of extending the state-of-the-art SF-IDF method for news recommendation, in order to additionally take into account semantic relations between synsets, as well as named entities. The proposed recommendation method, Bing-SF-IDF+, SF-IDF, and several SS methods have been implemented in Ceryx, an extension to the Hermes news personalization service for news recommendation. Our evaluation on 100 financial news messages and 8 user profiles (queries) showed that Bing-SF-IDF+ outperforms the other methods for F_1-scores and kappa statistics.

The discussed recommenders are based on synsets from a semantic lexicon. However, such recommenders are dependent on the information available in such lexicons. Therefore, as future work, we would like to investigate a way to combine multiple semantic lexicons, or to create an expert system which gathers information and updates the known information in a semantic lexicon. Additionally, we would like to experiment with obtaining the semantic relationship weights by making use of other metaheuristic methods such as simulated annealing and ant colony optimization. Last, it would be worthwhile to investigate similar Bing-based named entity extensions to other recommendation methods, such as TF-IDF and SS.

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. We thank SARA Computing and Networking Services for their support in using the Lisa Computer Cluster.

6. REFERENCES

[1] Alias-i. LingPipe 4.1.0. From: http://alias-i.com/lingpipe, 2008.

[2] Bing. Bing API 2.0. http://www.bing.com/developers/s/APIBasics.html, 2012.

[3] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring Semantic Similarity between Words Using Web Search Engines. In *16th Int. Conf. on World Wide Web (WWW 2007)*, pages 757–766. ACM, 2007.

[4] G. Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. In C. Chiarcos, R. E. de Castilho, and M. Stede, editors, *Biennial GSCL Conf. 2009 (GSCL 2009)*, pages 31–40. Gunter Narr Verlag Tübingen, 2009.

[5] M. Capelle, M. Moerland, F. Frasincar, and F. Hogenboom. Semantics-Based News Recommendation. In *2nd Int. Conf. on Web Intelligence, Mining and Semantics (WIMS 2012)*. ACM, 2012.

[6] R. Cilibrasi and P. M. B. Vitányi. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[7] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[8] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 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)*, pages 168–175. Association for Computational Linguistics, 2002.

[9] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

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

[11] F. Getahun, J. Tekli, R. Chbeir, M. Viviani, and K. Yetongnon. Relating RSS News/Items. In *9th Int. Conf. on Web Engineering (ICWE 2009)*, pages 442–452. Springer-Verlag, 2009.

[12] I. R. Group. WordWideWebSize.com. `http://www.worldwidewebsize.com`, 2012.

[13] A. S. Jensen and N. S. Boss. Textual Similarity: Comparing Texts in Order to Discover How Closely They Discuss the Same Topics. Bachelor's Thesis, Technical University of Denmark, 2008.

[14] J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *10th Int. Conf. on Research in Computational Linguistics (ROCLING 1997)*, pages 19–33, 1997.

[15] C. Leacock and M. Chodorow. *WordNet: An Electronic Lexical Database*, chapter Combining Local Context and WordNet Similarity for Word Sense Identification, pages 265–283. MIT Press, 1998.

[16] D. Lin. An Information-Theoretic Definition of Similarity. In *15th Int. Conf. on Machine Learning (ICML 1998)*, pages 296–304. Morgan Kaufmann, 1998.

[17] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *14th Int. Joint Conf. on Artificial Intelligence (IJCAI 1995)*, pages 448–453. Morgan Kaufmann, 1995.

[18] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Human Language Technology Conf. of the North American Chapter of the Association for Computational Linguistics (HLTNAACL 2003)*, pages 252–259, 2003.

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