News Recommendation using Semantics with the Bing-SF-IDF Approach

Frederik Hogenboom, Michel Capelle, and Marnix Moerland

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

fhogenboom@ese.eur.nl,
{michelcapelle, marnix.moerland}@gmail.com

Abstract. Traditionally, content-based news recommendation is performed by means of the cosine similarity and the TF-IDF weighting scheme for terms occurring in news messages and user profiles. Semantics-driven variants like SF-IDF additionally take into account term meaning by exploiting synsets from semantic lexicons. However, 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 employing named entity similarities using Bing page counts. Our proposed method, Bing-SF-IDF, outperforms TF-IDF and its semantics-driven variants in terms of F_1-scores and kappa statistics.

1 Introduction

The Web is an increasingly important source of information, which is mostly posted in the form of news. However, today's users are confronted with an overload of information, and hence, many recommendation methods have been developed that aid in filtering and structuring information based on user preferences or characteristics (captured in user profiles). Traditionally, such content-based recommender systems are based on term frequencies. A commonly used measure is the Term Frequency – Inverse Document Frequency (TF-IDF) [16]. When employing user profiles, 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. For this, TF-IDF weights are computed on every term within a document.

TF-IDF-based systems do not consider the text semantics, which could be added by using Web ontologies. A drawback of ontologies is that they are domain dependent, and hence require continuous maintenance. Alternatively, one could employ synonym sets (synsets) from general semantic lexicons (e.g., WordNet [8]). In earlier work [5], we introduced the Synset Frequency – Inverse Document Frequency (SF-IDF) measure, operating on WordNet synsets instead of terms. We evaluated SF-IDF against TF-IDF and against an other semantics-based alternative, Semantic Similarity (SS), and demonstrated the benefits of considering synsets.

Generally, news articles are linked to many named entities. These could provide crucial information when constructing user profiles, yet when performing synset-based recommendation, they are often not considered. One could enhance existing semantics-based recommendation methods like SF-IDF by employing similarities based on page counts gathered by Web search engines, such as Google or Bing, thus avoiding the use of domain dependent ontologies. Our current endeavors contribute to the state-of-the-art by extending SF-IDF by additionally considering 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 semantics-driven Bing-SF-IDF news recommender in Section 3. Subsequently, we evaluate our performance in Section 4. Last, we draw conclusions and provide some directions for future work in Section 5.

2 Preliminaries

In the field of recommender systems, many different profile-based recommender systems have been developed. Their main difference lies in the implemented similarity measures for calculating news item and user profile similarities.

2.1 TF-IDF

The de facto standard of similarity measures found in literature is TF-IDF, combined with cosine similarities [14]. The TF-IDF method consists of two parts, i.e., term frequency $tf(t,d)$ and inverse document frequency $idf(t,d)$. It operates on terms T in documents D and 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 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 to compare, and $d : t \in d$ represents the amount of documents containing term t. Next, $T(t,d)$ is obtained by multiplying $tf(t,d)$ and $idf(t,d)$:

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

Subsequently, 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. Then, we obtain the similarity between a set of terms from news item d_u and user profile d_r using the cosine similarity measure:

$$sim_T(d_u, d_r) = \frac{A(d_u) \cdot A(d_r)}{||A(d_u)|| \times ||A(d_r)||} \ . \tag{4}$$

After every unread news document has been assigned a similarity value (with respect to the user profile), all unread news items with similarities higher than a cut-off value are recommended to the user.

2.2 SF-IDF

One of the main drawbacks of TF-IDF is that semantics are not considered. This causes synonyms to be mistakenly counted as separate terms, and homonyms as one term. Therefore, semantics-based similarity measures have been proposed, such as the Synset Frequency – Inverse Document Frequency (SF-IDF) [5], i.e., a TF-IDF variant which makes use of synonym sets (synsets) from a semantic lexicon (e.g., WordNet [8]) instead of terms. These synsets are obtained after a word sense disambiguation procedure. When replacing term t by synset s, the SF-IDF formulas are:

$$sf(s, d) = \frac{n_{s,d}}{\sum_k n_{k,d}} \ , \tag{5}$$

$$idf(s, d) = \log \frac{|D|}{|\{d : s \in d\}|} \ , \tag{6}$$

$$S(s, d) = sf(s, d) \times idf(s, d) \ , \tag{7}$$

$$sim_S(d_u, d_r) = \frac{A(d_u) \cdot A(d_r)}{||A(d_u)|| \times ||A(d_r)||} \ . \tag{8}$$

News item recommendation is subsequently performed in a similar manner as for the TF-IDF method, using the cosine similarity measure and a cut-off value.

2.3 Semantic Similarity

The Semantic Similarity (SS) method [5] compares synsets from unread news items with those of the user profile by employing pairs between the elements of the two sets with a common part-of-speech (i.e., they elements share a linguistic category such as a verb or a noun). We define $V = U \times R$, i.e., the Cartesian product of U and R containing all possible combinations of synsets from the unread news item d_u, referred to as U, and the union of synsets from the user profile d_r, denoted by R. Here u_k and r_l denote synsets from the unread news item and user profile, and k and l are the number of available synsets. A subset of V containing all the combinations with common parts-of-speech is defined as $W = \{(u, r) \in V : POS(u) = POS(r)\}$, where $POS(u)$ and $POS(r)$ describe 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. This rank measures the semantic distance between synsets u and r when represented as nodes in a hierarchy of 'is-a' relationships:

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

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

The similarity rank $sim(u,r)$ can be computed in various ways. Some make use of the information content (the negative logarithm of the sum of all probabilities of all the words in the synset). For instance, the Jiang & Conrath [11] measure uses the information content of both the synsets and the lowest common subsumer, while Lin's measure [13] makes use of the logarithms of the chances of appearance of both nodes and the lowest common subsumer. Resnik's measure [15] on the other hand maximizes the information content of the lowest common subsumer of the two nodes. The Leacock & Chodorow [12] and Wu & Palmer [18] measures on the other hand make use of the path length between the nodes. The path length is either the shortest path between the two nodes (Leacock & Chodorow) or the depth from a node to the top node (Wu & Palmer).

2.4 Enhancements

Recent studies have shown that 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]. Therefore, one could enhance existing semantics-based recommendation methods, such as SF-IDF and SS, by (additionally) employing similarities based on page counts gathered by Web search engines, such as Google or Bing.

3 Bing-SF-IDF Recommendation

Like most semantics-based news recommendation methods, Bing-SF-IDF operates on a user profile, containing all currently read news items. For each unread news item, a similarity score between the news article and the user profile is computed. In case a similarity score exceeds a predefined cut-off value, the corresponding news item is recommended. In essence, the Bing-SF-IDF similarity score is a weighted average of two similarity scores. The Bing component expresses similarities between named entities, whereas SF-IDF measures the similarities between synsets.

The Bing similarity score takes into account the named entities that do not occur in a semantic lexicon, by deriving them from news articles through a named entity recognizer. For this, we consider an unread news item d_u and user profile d_r, which can be described respectively using two sets of named entities, i.e.,

$U = \{u_1, u_2, \ldots, u_k\}$, $R = \{r_1, r_2, \ldots, r_l\}$. Here, u_i (where $1 \leq i \leq k$) represents a named entity in the unread news item U, r_j (where $1 \leq j \leq l$) describes 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 $V = U \times R$ containing all possible pairs of named entities from the unread news item d_u and the user profile d_r. Then, between-pair similarity is measured using search engine page counts, i.e., the number of Web pages found by the Bing Web search engine containing 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], which we define as:

$$sim_{PMI}(u, r) = \log \frac{\frac{c(u,r)}{N}}{\frac{c(u)}{N} \times \frac{c(r)}{N}} , \tag{10}$$

where $c(u, r)$ denotes the page count for the pair (u, r) of named entities. Moreover, $c(u)$ 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 is the amount of Web pages indexed by Bing (∼15 billion).

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

$$sim_B(V) = \frac{\sum\limits_{(u,r) \in V} sim_{PMI}(u, r)}{|V|} . \tag{11}$$

Next, we can combine the Bing score sim_B and SF-IDF similarity score sim_S for every unread news item d_u and user profile d_r. We employ min-max normalization between 0 and 1 on both sets of similarity scores, and subsequently take the weighted average of these scores:

$$\overline{sim}_B(d_u, d_r) = \frac{sim_B(d_u, d_r) - \min\limits_u sim_B(d_u, d_r)}{\max\limits_u sim_B(d_u, d_r) - \min\limits_u sim_B(d_u, d_r)} , \tag{12}$$

$$\overline{sim}_S(d_u, d_r) = \frac{sim_S(d_u, d_r) - \min\limits_u sim_S(d_u, d_r)}{\max\limits_u sim_S(d_u, d_r) - \min\limits_u sim_S(d_u, d_r)} , \tag{13}$$

$$sim_{BS}(d_u, d_r) = \alpha \times \overline{sim}_B(d_u, d_r) + (1 - \alpha) \times \overline{sim}_S(d_u, d_r) . \tag{14}$$

Here, weight α is optimized during testing 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.

Our framework is implemented as an extension to the Ceryx [5] plugin of the Hermes News Portal (HNP) [9]. The Java-based HNP operates based on user profiles and processes news items from RSS feeds, while making use of an expert-created OWL domain ontology. News items are classified using the GATE natural language processing software [7] and the WordNet [8] semantic

lexicon. The semantics-based methods make use of the Stanford Log-Linear Part-of-Speech Tagger [17], Lesk Word Sense Disambiguation [10], and the Alias-i's LingPipe 4.1.0 [1] Named Entity Recognizer. Page counts are collected with the Bing API 2.0 [2].

4 Evaluation

For our evaluation of the performance of Bing-SF-IDF compared against its alternatives, we collected 100 news articles from a Reuters news feed on technology companies. Although the set of articles appears to be rather small, the limited number of articles is motivated by the fact that subsequent annotation steps require a lot of effort from the annotators, and hence for now we suffice with 100 annotated news articles. In our experiments, three experts classified these articles based on 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% The overall agreement was on average 90%. 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 our proposed recommendation method, we compare the performance to the performance of the TF-IDF [16], SF-IDF [5], and five SS recommendation methods [11–13, 15, 18] in terms of F_1 (i.e., the harmonic mean of precision and recall scores) and kappa statistics [6] (measuring whether the proposed classification is better than a random guess), which are the norm in this context. We additionally report on accuracy, precision, recall, sensitivity, and specificity. Moreover, 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 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 150, a mutation probability of 0.1, elitism of 50, and a maximum number of 25 generations.

4.2 Experimental Results

The average performance for each recommender is displayed in Table 1. From this, we can conclude that Bing-SF-IDF outperforms all recommenders, and demonstrates to be a substantial improvement over SF-IDF method which served as a basis for our proposed extension. Also, the Jiang & Conrath recommender shows good overall performance. The graphs in Figures 1(a) and 1(b) provide a closer look into the F_1-scores and kappa statistics for all cut-off values and 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. For the

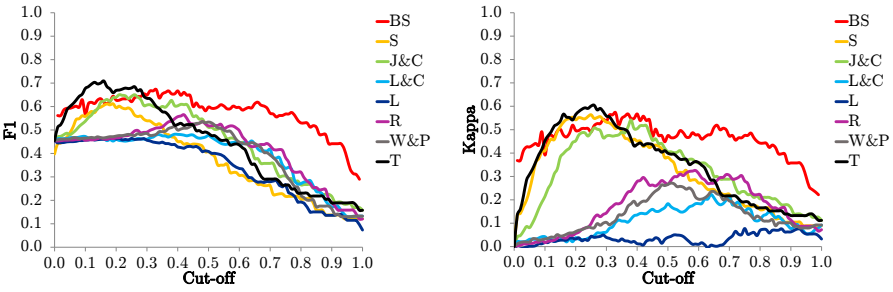
Table 1. Average test results for Bing-SF-IDF (BS), SF-IDF (S), Jiang & Conrath (J&C), Leacock & Chodorow (L&C), Lin (L), Resnik (R), Wu & Palmer (W&P), and TF-IDF (T).

	Acc.	Prec.	Rec.	F_1	Sens.	Spec.	Kappa
BS	0.81	0.69	0.53	0.58	0.53	0.91	0.46
S	0.65	0.68	0.43	0.37	0.37	0.76	0.32
J&C	0.72	0.73	0.48	0.45	0.47	0.82	0.31
L&C	0.55	0.44	0.58	0.39	0.58	0.54	0.11
L	0.51	0.38	0.53	0.34	0.53	0.51	0.03
R	0.60	0.55	0.57	0.42	0.57	0.61	0.17
W&P	0.57	0.46	0.59	0.40	0.59	0.55	0.13
T	0.75	0.83	0.44	0.45	0.43	0.88	0.34

other recommendation methods, performances do not differ a lot within the high range of cut-off values. For low cut-off values, on the other hand, TF-IDF performs best.

The kappa statistic, which is plotted in Figure 1(b), measures whether the proposed classifications are better than random guessing. Positive scores (better than random) are preferred over negative scores (worse than random) or scores of exactly 0 (same as random). The Bing-SF-IDF recommender scores a higher kappa statistic than the other recommenders for high cut-off values, indicating that the Bing-SF-IDF recommender seems to have more classification power than the others. Also, TF-IDF and the semantics-based SF-IDF and Jiang & Conrath SS methods show good performance.

The statistical significance of the results is assessed in Tables 2 and 3. We can deduce that Bing-SF-IDF significantly outperforms all other approaches in terms of F_1-scores and kappa statistics. Also, TF-IDF performs well, as it significantly



(a) F_1-scores. (b) Kappa statistics.

Fig. 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.

Table 2. 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 3. 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	

outperforms SF-IDF and all SS methods when it comes to kappa statistics, yet it does not significantly outperform the Jiang & Conrath SS method in terms of F_1. SF-IDF merely outperforms one recommender in terms of F_1 (Lin SS), but seems to do better in terms of kappa statistics. The worst performing recommendation method overall is Lin SS, which is significantly outperformed by all methods on both measures.

Last, an evaluation of the optimized α-values for all cut-off values leads to various insights. For Bing-SF-IDF, scores are weighted using an average optimized α of 0.52 (with a standard deviation of 0.29). Hence, a substantial weight is given to both the Bing similarities and the synsets, underlining the importance of both proposed extensions. Worth noting is that the higher the cut-off, the higher the value of α. This indicates that Bing similarities become more important when a high precision is required. For lower cut-off values, i.e., when the focus is more on high recall, named entities seem less relevant and synsets alone are more than adequate for recommending news items.

5 Conclusions

In general, news recommendation is performed using the cosine similarity and the term-based TF-IDF weighting scheme. However, semantics-driven methods, which take into account term meaning, are able to handle news information in a better way. Such methods exploit semantic lexicon synsets and the cosine similarity (SF-IDF) or make use of semantic similarities (SS). However, they do not take into account named entities, which are usually not present in semantic lexicons.

Hence, we extended the state-of-the-art SF-IDF recommendation method by also taking into account named entities using Bing page counts. Our proposed method, Bing-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 Bing-SF-IDF significantly outperforms TF-IDF as well as other semantic methods with respect to F_1-scores and kappa statistics.

The discussed recommenders are based on synsets from a single semantic lexicon. However, this still creates a dependency on the information available in such lexicons. Therefore, as future work, we would like to investigate a way to combine multiple semantic lexicons. Moreover, it would be worthwhile to explore the possibilities of employing semantic relations from semantic lexicons, or to perform additional analysis on similar Bing-based named entity extensions to other recommendation methods, such as 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.

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. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring Semantic Similarity between Words Using Web Search Engines. In: 16th Int. Conf. on World Wide Web (WWW 2007). pp. 757–766. ACM (2007)
4. Bouma, G.: Normalized (Pointwise) Mutual Information in Collocation Extraction. In: Chiarcos, C., de Castilho, R.E., Stede, M. (eds.) Biennial GSCL Conf. 2009 (GSCL 2009). pp. 31–40. Gunter Narr Verlag Tübingen (2009)
5. Capelle, M., Moerland, M., Frasincar, F., Hogenboom, F.: Semantics-Based News Recommendation. In: 2nd Int. Conf. on Web Intelligence, Mining and Semantics (WIMS 2012). ACM (2012)
6. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1), 37–46 (1960)

7. 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)
8. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
9. Frasincar, F., Borsje, J., Levering, L.: A Semantic Web-Based Approach for Building Personalized News Services. Int. J. E-Business Research 5(3), 35–53 (2009)
10. 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)
11. Jiang, J.J., Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: 10th Int. Conf. on Research in Computational Linguistics (ROCLING 1997). pp. 19–33 (1997)
12. 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)
13. Lin, D.: An Information-Theoretic Definition of Similarity. In: 15th Int. Conf. on Machine Learning (ICML 1998). pp. 296–304. Morgan Kaufmann (1998)
14. Marie-Francine Moens: Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer (2006)
15. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: 14th Int. Joint Conf. on Artificial Intelligence (IJCAI 1995). pp. 448–453. Morgan Kaufmann (1995)
16. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)
17. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: 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). pp. 252–259 (2003)
18. 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)