



News Personalization using the CF-IDF Semantic Recommender

Frank Goossen

frank.goossen@xs4all.nl

Wouter IJntema

wouterijntema@gmail.com

Flavius Frasincar

frasincar@ese.eur.nl

Frederik Hogenboom

fhogenboom@ese.eur.nl

Uzay Kaymak

kaymak@ese.eur.nl

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

May 25, 2011

International Conference on Web Intelligence, Mining, and Semantics (WIMS 2011)



Introduction (1)

- Recommender systems help users to plough through a massive and increasing amount of information
- Recommender systems:
 - Content-based
 - Collaborative filtering
 - Hybrid
- Content-based systems are often term-based
- Common measure: Term Frequency – Inverse Document Frequency (**TF-IDF**) as proposed by Salton and Buckley [1988]

Introduction (2)

- TF-IDF steps:
 - Filter stop words from document
 - Stem remaining words to their roots
 - Calculate term frequency (i.e., the importance of a term or word within a document)
 - Calculate inverse document frequency (i.e., the inverse of the general importance of a term in a set of documents)
 - Multiply term frequency with the inverse document frequency
- TF-IDF performance tends to decrease as documents get larger



Introduction (3)

- The Semantic Web offers new possibilities
- Utilizing concepts instead of terms:
 - Reduces noise caused by non-meaningful terms
 - Yields less terms to evaluate
 - Allows for semantic features, e.g., synonyms
- Therefore, we propose Concept Frequency – Inverse Document Frequency (**CF-IDF**)
- CF-IDF is implemented in **Athena** (an extension for **Hermes** [Frasincar et al., 2009], a news processing framework)
- Results are evaluated in comparison with TF-IDF

Introduction (4)

- Earlier work has been done:
 - CF-IDF-like methods: Baziz et al. [2005], Yan and Li [2007]
 - Frameworks: OntoSeek [Guarino et al., 1999], Quickstep [Middleton et al., 2004], News@hand [Cantador et al., 2008]
- Although some work shows overlap:
 - Methods are not thoroughly compared with TF-IDF
 - Often, WSD and synonym handling is lacking





Outline

- TF-IDF
- CF-IDF
- Recommendations
- Implementation:
 - Hermes
 - Athena
- Evaluation
- Conclusions

TF-IDF

- Term Frequency: the occurrence of a term t_i in a document d_j , i.e.,

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- Inverse Document Frequency: the occurrence of a term t_i in a set of documents D , i.e.,

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

- And hence

$$tf - idf_{i,j} = tf_{i,j} \times idf_i$$

CF-IDF

- Concept Frequency: the occurrence of a concept c_i in a document d_j , i.e.,

$$cf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- Inverse Document Frequency: the occurrence of a concept c_i in a set of documents D , i.e.,

$$idf_i = \log \frac{|D|}{|\{j : c_i \in d_j\}|}$$

- And hence

$$cf - idf_{i,j} = cf_{i,j} \times idf_i$$

Recommendations

- Ontology contains a set of concepts and relations
- User profile consists of (a subset of) these concepts and relations
- Each concept and relation is associated with all news articles
- Each article is represented as:
 - TF-IDF: a set containing all terms
 - CF-IDF: a set containing all concepts
- Then, for each article, weights are calculated
- Weights of a new article are compared to the user profile using cosine similarity



Implementation: Hermes

- Hermes framework is utilized for building a news personalization service
- Its implementation Hermes News Portal (**HNP**):
 - Is ontology-based
 - Is programmed in Java
 - Uses OWL / SPARQL / Jena / GATE / WordNet
- Input: RSS feeds of news items
- Internal processing:
 - Classification
 - News querying
- Output: news items

Implementation: Athena (1)

- Athena is a plug-in for HNP
- Main focus is on recommendation support
- User profiles are constructed
- TF-IDF (using a stemmer as proposed in [Krovetz, 1993]) and CF-IDF recommendation calculations can be performed



Implementation: Athena (2)

- Interface:
 - News browser
 - Recommendations
 - Evaluation



Implementation: Athena (3)



The screenshot displays the Hermes News Portal v1.1 interface. At the top, there are navigation tabs: Home, Original graph, Search graph, Results, Recommendations, Users, and Evaluation. Below these, there are sub-tabs: All News Items, Recommendations, and Test Results. The main content area shows search results for 'CF-IDF Recommender' with 10 items. The results are listed as follows:

- Item #1:** **Reviewing Bing from Back to Front** (2009-06-16 08:55:56 Rank: 100). The text discusses a news analysis about Bing's search program and its privacy features.
- Item #2:** **New Google tool targets Microsoft business users** (2009-06-09 11:19:45 Rank: 93). The text reports on Google's new software designed to make it easier for businesses using Microsoft products.
- Item #3:** **Opera Unite Debuts to Challenge Google, Mozilla in Web Services** (2009-06-16 06:21:00 Rank: 70). The text describes Opera's new platform, Unite, which aims to challenge Google and Mozilla in the web services market.
- Item #4:** **Rajeev Motwani, Guide in the Creation of Google, Dies at 47** (2009-06-11 06:50:30 Rank: 69). The text mentions the death of Professor Motwani, a mentor to many Silicon Valley start-ups.

On the right side of the interface, there is a **Tag Cloud** with the following tags: AMZN, BecomesCEO, Company, Deal, **GOOG**, MSFT, New York Stock Exchange, NewCompetitor, Revenue, Sells, Software, and United States.

Implementation: Athena (4)



The screenshot shows the 'Hermes News Portal v1.1' interface. The 'Test Results' tab is active, displaying performance metrics for two recommender algorithms. The 'CF-IDF Recommender' shows 8.0 True Positives, 30.0 True Negatives, 93% Accuracy, 100% Precision, 0.0 False Positives, 3.0 False Negatives, 73% Sensitivity, and 100% Specificity. The 'TF-IDF Recommender' shows 2.0 True Positives, 30.0 True Negatives, 78% Accuracy, 100% Precision, 0.0 False Positives, 9.0 False Negatives, 18% Sensitivity, and 100% Specificity. The interface also includes a 'Random Seed' field set to 0, a 'Cut-Off' field set to 0.4, a 'Done' status bar, and buttons for 'Reset Results' and 'Save Results'. Summary statistics at the bottom indicate 100 rated news items and a validation set size of 41.

Recommender	True Positives	True Negatives	Accuracy	Precision	False Positives	False Negatives	Sensitivity	Specificity
CF-IDF Recommender	8.0	30.0	93%	100%	0.0	3.0	73%	100%
TF-IDF Recommender	2.0	30.0	78%	100%	0.0	9.0	18%	100%

Testfile: test.xml
Iteration: 25
Number of tests: 25
 Cut-Off Testing
Reset Results
Save Results
Rated news items: 100
Size of validation set: 41

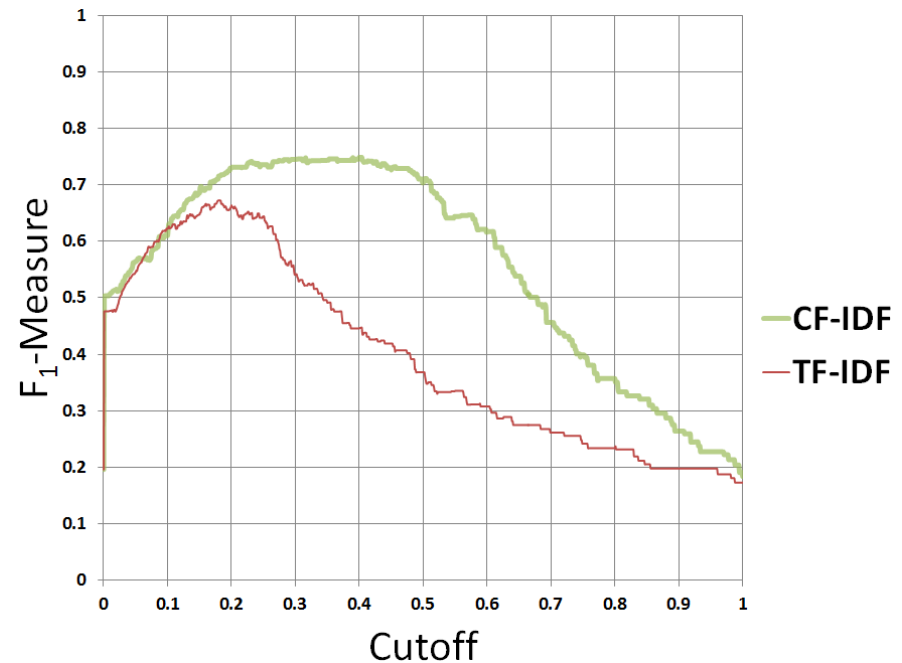
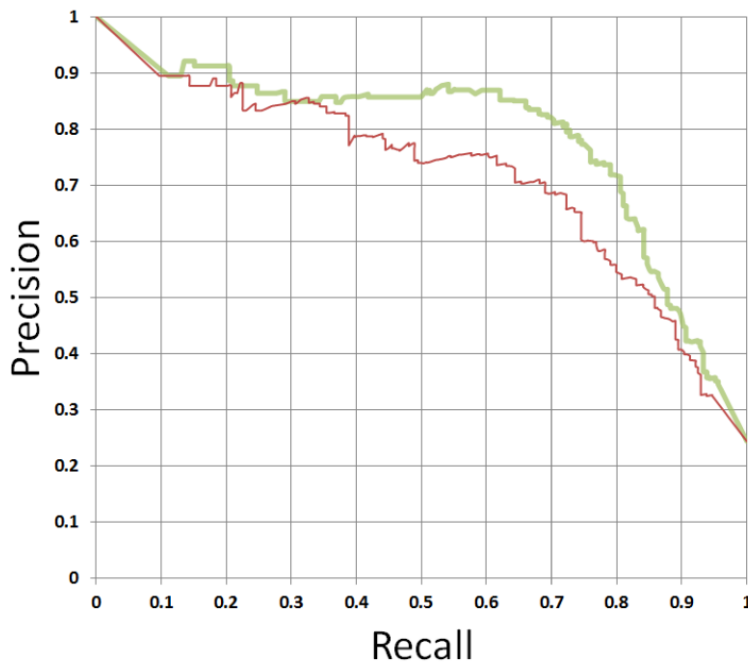
Evaluation (1)

- Experiment:
 - We let 19 participants evaluate 100 news items
 - User profile: all articles that are related to Microsoft, its products, and its competitors
 - Athena computes TF-IDF and CF-IDF and determines interestingness using several cutoff values
 - Measurements:
 - Accuracy
 - Precision
 - Recall
 - Specificity
 - F_1 -measure
 - Kappa statistic
 - Receiver Operating Characteristic (ROC) curves
 - t -tests for determining significance

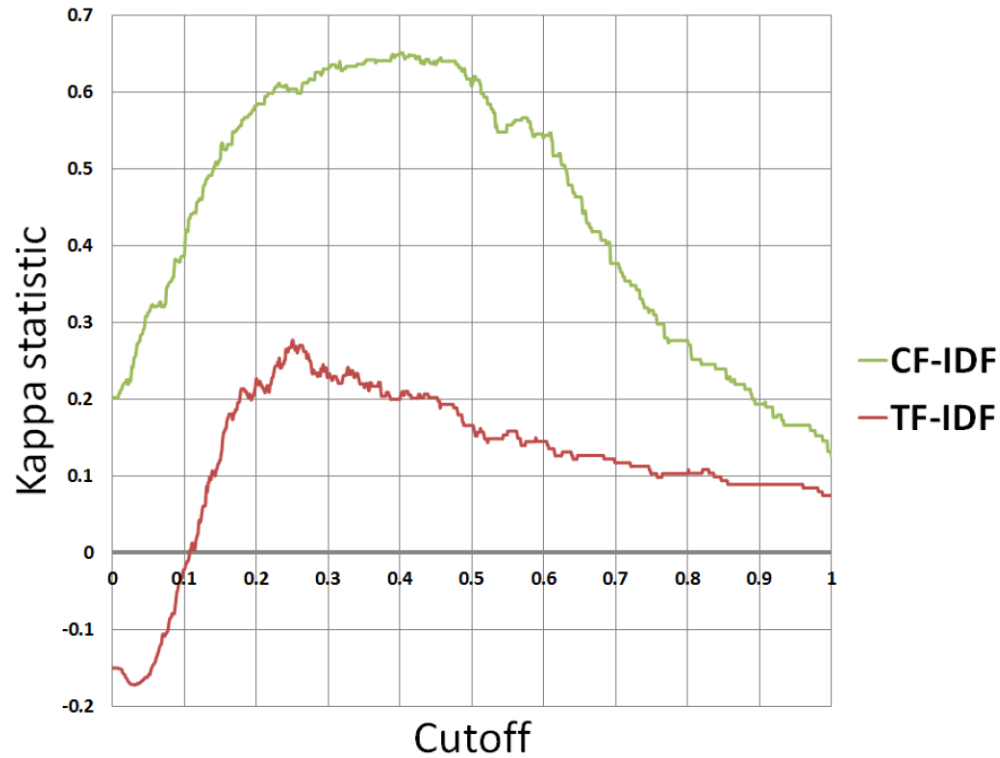


Evaluation (2)

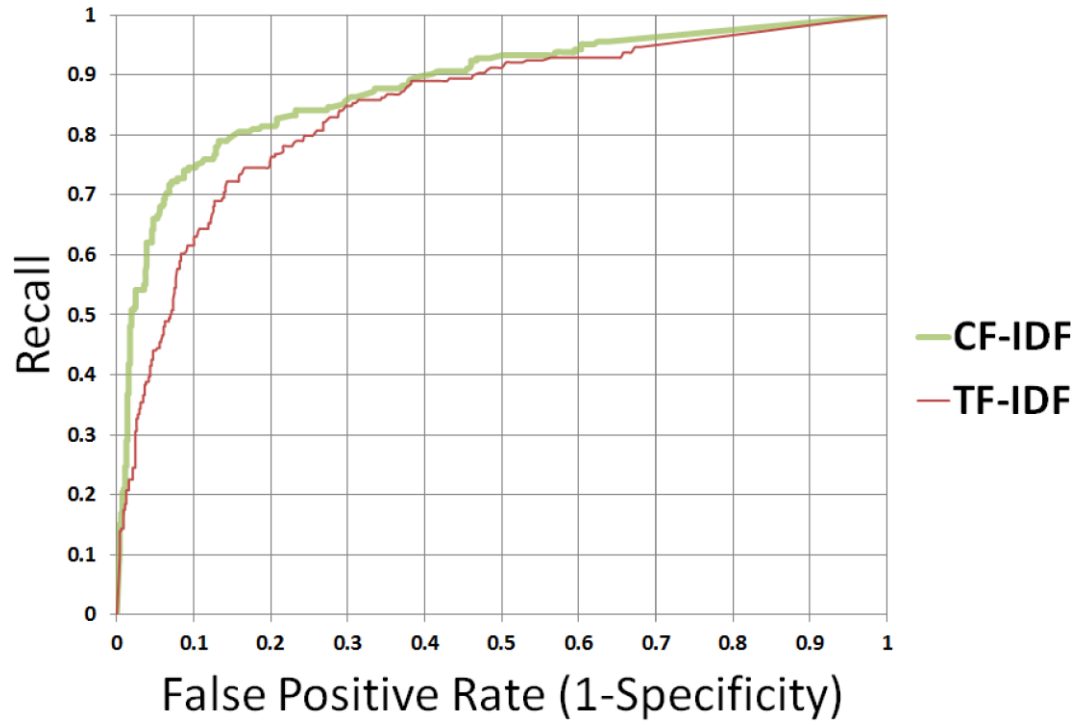
- Results:
 - CF-IDF performs significantly better than TF-IDF for accuracy (+4.7%), recall (+24.4%), and F_1 (+21.9%) for threshold 0.5
 - Precision and specificity are not significantly different



Evaluation (3)



Evaluation (4)



Conclusions

- CF-IDF outperforms TF-IDF significantly for many measures: accuracy, recall, F_1 , Kappa, and ROC (AUC)
- Hence, using key concepts and semantics instead of analyzing all terms could be beneficial for recommender systems
- Future work:
 - Use different stemmers for TF-IDF
 - Investigate and compare with TF-IDF variants that account for some limitations (e.g., Okapi BM25)
 - Implement various concept relationship types

Questions



References (1)

- Baziz, M., Boughanem, M., Traboulsi, S.: A Concept-Based Approach for Indexing Documents in IR. In: Actes du XXIIIème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2005). pp. 489-504. HERMES Science Publications (2005)
- Cantador, I., Bellogín, A., Castells, P.: News@hand: A Semantic Web Approach to Recommending News. In: 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2008). pp. 279-283. Springer-Verlag, Berlin, Heidelberg (2008)



References (2)

- Frasinca, 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)
- Guarino, N., Masolo, C., Vetere, G.: *OntoSeek: Content-Based Access to the Web*. *IEEE Intelligent Systems* 14(3), 70-80 (1999)
- Krovetz, R.: *Viewing Morphology as an Inference Process*. In: *26th ACM Conference on Research and Development in Information Retrieval (SIGIR 1993)*. pp. 191-202. ACM (1993)



References (3)

- Middleton, S.E., Roure, D.D., Shadbolt, N.R.: Ontology-Based Recommender Systems. In: Handbook on Ontologies, pp. 577-498. International Handbooks on Information Systems, Springer (2004)
- Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513-523 (1988)
- Yan, L., Li, C.: A Novel Semantic-based Text Representation Method for Improving Text Clustering. In: 3rd Indian International Conference on Artificial Intelligence (IICAI 2007). pp. 1738-1750 (2007)

