A Comparison Study for Novelty Control Mechanisms Applied to Web News Stories

Arnout Verheij, Allard Kleijn, Flavius Frasincar, and Frederik Hogenboom

Econometric Institute, Erasmus University Rotterdam
PO Box 1738, NL-3000 DR, Rotterdam, the Netherlands
Email: verheij.a@gmail.com, allardkleijn@hotmail.com, {frasincar, fhogenboom}@ese.eur.nl

***Abstract*—In this paper we evaluate several novelty control mechanisms for ranking Web news articles depicting a story. These mechanisms rank individual news items based on a novelty measure using as context the items which have been previously reviewed. An evaluation within the Hermes news personalization framework is performed for pairwise and non-pairwise novelty control mechanisms based on various distance measures and vector-based news representations. On average, the most effective distance measure is Kullback-Leibler and the best performing news representation vector uses named entities.**

***Keywords*-novelty control; ranking; Web news**

I. Introduction

The Web is a great source of information, affecting many of our daily information-intensive activities. Among others, this information comes in the form of news items. However, the user is confronted with an overload of information, and hence, many methods have been developed to filter and structure this information. Often, these approaches are based on queries made by the user to represent his interests. One framework using this approach is Hermes [1]–[3]. Hermes is a framework for personalizing news based on Semantic Web technologies, which uses concepts from a knowledge base instead of simple keywords for annotating and querying. Unfortunately, this approach retrieves a lot of relevant articles which offer no new information when compared to the story depicted by previously reviewed articles.

Once a user finds a story which is relevant to his interests, he might want to keep following this story. Story-based news representation combined with novelty control could allow the user to identify and keep track of developing stories. Novelty control is a mechanism which sorts news items based on their novelty compared to the seed item and previously browsed items [4]. The seed item usually is the first item from the story. Similar to grouping of news in stories, novelty control makes use of distance measures to determine the similarity of two documents. News items which are dissimilar to other documents, but which belong to the same story might indicate that the storyline is developing and new information is released. Such documents should receive a higher novelty score than others.

Most methods developed for these purposes use all words from the documents as dimensions in a vector-based news representation [4]. Common algorithms use language models based on these words and compute the Kullback-Leibler (KL) divergence [5]. In our approach we use all words as well as the named entities as vector dimension. We assume that named entities carry a large part of the story information contained in the news item. It is possible that using all words also includes a lot of noise, words that do not depict the story of interest. We examine the use of both named entities and all words in combination with several distance measures. The main contribution of this paper is our extensive evaluation of different novelty control mechanisms using various distance measures and vector-based news representations, which to our knowledge has not been done before.

The rest of this paper is organized as follows. Section II describes related work on novelty control. The algorithms of our framework are devised in Section III. Then, Section IV discusses the implementation of our algorithms in the Hermes News Portal, the implementation of the Hermes framework. Section V evaluates the investigated novelty control solutions and Section VI concludes this paper.

II. Related Work

This section gives a brief overview of related research in the field of novelty control. In the first subsection we discuss general mechanisms of novelty control. In the remaining three sections we describe existing document distance measures along with their specific data representations.

A. Novelty Control

In 1998, Carbonell and Goldstein motivated a need for novelty control in addition to the traditional ranking of documents which is solely based on the user's query [6]. They devised the Maximal Marginal Relevance (MMR) method which gives a document a high marginal relevance if it is both relevant to the query and contains minimal similarity to previously selected documents. The original discussion on MMR lacks a description of possible similarity measures, as it only indicates that the metrics of traditional IR could be used. Novelty control can use both symmetric and asymmetric distance measures, because the order of news might matter as there is a history of browsed news.

Novelty control involves sorting the news based on a novelty score of news items compared to already browsed items. This novelty score usually consists of a traditional distance metric based on a certain document representation. The distance metric can be used either pairwise or non-pairwise [7]. Pairwise usage involves comparing a document to all previously browsed documents and determining novelty scores by computing the similarity to the most similar document, i.e.,

$$Nov(d_i|d_1,\ldots,d_m) = \min_{1\leq j\leq m} Dist(DR_i|DR_j) \ , \qquad (1)$$

where

$$\begin{aligned} Nov &= \text{Novelty score} \\ d_i &= \text{Document } i \\ j &= \text{Index of previously browsed documents} \\ m &= \text{Number of previously browsed documents} \\ Dist &= \text{Distance metric} \\ DR_i &= \text{Document representation of document } i \end{aligned}$$

With non-pairwise (or aggregated) usage, the data representations from all previously browsed documents are aggregated, and a document is compared to this aggregated vector. This aggregated distance metric is defined as:

$$Nov(d_i|d_1,\ldots,d_m) \quad = \quad Dist(DR_i|DR_u) \ , \qquad (2)$$

where

$$DR_u = \bigcup_{j=1}^{m} DR_j$$

B. Language Models

A common method to measure the similarity between two documents is using language models [8]. Language models are models that define a probability mechanism for generating language. They compute a probability P of picking term T randomly from document D. This probability is called $P(T|D)$. A language model Θ from document D can be defined as the vector of probabilities $(P(T_1|D), P(T_2|D), \ldots, P(T_n|D))$ where n is the number of different terms in the set of documents. Please note that we focus here solely on unigrams (in contrast to bigrams, trigrams, etc.) when we refer to language models.

Divergence metrics compute the distance between two language models Θ_1 and Θ_2, indicating the dissimilarity. A commonly used divergence metric is Kullback-Leibler (KL) divergence [5]. KL divergence is asymmetric and therefore only applicable to novelty control and not to story detection. It is defined as:

$$KL(\Theta_1||\Theta_2) = \sum_{i=1}^{n} \Theta_1(i) log \frac{\Theta_1(i)}{\Theta_2(i)} \ , \qquad (3)$$

where

$$\begin{aligned} \Theta_x &= \text{Language model based on document } x \\ \Theta_x(i) &= \text{Probability of term } T_i \text{ in document } x \end{aligned}$$

Another divergence measure is Jensen-Shannon (JS) [9] divergence. JS divergence is a symmetric and smoothed variant of the KL divergence and is defined as follows:

$$JS(\Theta_1||\Theta_2) = \frac{1}{2}KL(\Theta_1||M) + \frac{1}{2}KL(\Theta_2||M) \ , \qquad (4)$$

where

$$M = \frac{1}{2}(\Theta_1 + \Theta_2) \ .$$

Because we only consider terms which are in either of the documents, JS divergence will always give finite values as result. A problem with the KL divergence in language models is that some probabilities might be zero, making the denominator zero. This problem can be solved by means of a smoothing method, i.e., linear interpolation smoothing (also called Jelinek-Mercer smoothing) [10], smoothing document weights for KL divergence against the set of all documents [8]. The smoothing calculations are shown in (5).

$$\Theta_k(i) \quad = \quad \lambda * \Theta_k(i) + (1-\lambda) * \Theta_{1,\ldots,n}(i) \ , \qquad (5)$$

where

$$\begin{aligned} k &= \text{1 or 2} \\ \lambda &= \text{Unknown value between 0 and 1 which} \\ &\quad\ \text{should be tuned to optimize performance} \\ \Theta_{1,\ldots,n}(i) &= \text{Probability of term } T_i \text{ in aggregated} \\ &\quad\ \text{documents } 1, \ldots, n \end{aligned}$$

C. Word Count Models

Other frequently used methods to compute the distance between two documents are based on counting words in both documents. A straightforward method of calculating the distance between two documents is a new word count [11]. The score of the new document is equal to the number of words in it which have not appeared in the old one. Please note that this distance metric is asymmetric. This method might be used either pairwise or aggregated. The pairwise form of this method is referred to as set difference [12] and is considered to be more sophisticated [7].

D. Vector Space Model

Vector space models are commonly used for computing the distance between two documents, where documents are represented by vectors of weights. Some examples of data representations for the vector space model are Boolean weighting, a simple word count, and term frequency-inversed document frequency weighting (TF-IDF) [13], which has many variations. A popular distance metric for the vector space model is the cosine distance [14], which has proven to be the most successful metric to date for New Event Detection (NED) [15]. The cosine measure for novelty control could either use the pairwise comparison or the aggregate comparison [7]. The cosine distance is symmetric.

III. Novelty Control

An approach aimed at making it easier for the user to browse the news is by sorting the news based on their novelty or redundancy with respect to previously browsed items. Many news items about the same event are summarizing or repeating the same facts. These duplicate news items need to be filtered out. Also, it would be useful to rank news items which report many new facts about a story high in the result list. Users would be able to quickly inform themselves with new facts about the events they are interested in without having to read the same information again. In this section we present the investigated methods for novelty control.

Novelty control usually takes place within a single story, as novelty control without story-based news presentation is not interesting because two news items about completely different topics are usually novel with respect to each other. Traditional novelty control makes use of the same distance measures as a traditional ranking system with a query. However, browsing the news using a query requires the system to rank the news items which are very similar to the query high. In novelty control the browsed items are regarded as the query, however, in contrast to the traditional system, the items which are very dissimilar to this query should be ranked high. Novelty control can be implemented with both symmetric and asymmetric distance measures, which could be used either pairwise or aggregate.

In our approach we use the Kullback-Leibler divergence, Jensen-Shannon divergence, and the cosine similarity measure. KL and JS are used in combination with raw probabilities of terms (basic language models), while we use the cosine with TF-IDF and cosine normalization. Because novelty control uses distance metrics, we convert the cosine similarity measure as indicated in (6). All methods are evaluated both pairwise and aggregate, and both with the vector of all words and a vector of all named entities. News items which are similar to formerly read items should appear at the bottom of the result list.

$$Dist(d_i, d_j) = 1 - Sim(d_i, d_j) \ . \tag{6}$$

IV. Implementation

In this section we discuss the implementation of the novelty control algorithm in the Hermes News Portal (HNP) [1]–[3]. The HNP is a possible implementation of the Hermes framework. It provides the opportunity to employ the semantic features of the Hermes framework for news personalization.

The HNP is written in Java and uses OWL [16], recommended by W3C, as ontology language to store the news items. SPARQL [17], the query language recommended by W3C as the query language for RDF languages including OWL, is used to query this ontology. The Java library ARQ [18] is used to execute SPARQL queries.

In order to evaluate the novelty control methods, a suitable dataset is required. Therefore a wrapper is constructed to extract news items from a reliable source with an extensive offer of articles. The wrapper makes use of an array of existing packages in combination with custom code to extract the required information. First, using a package called TagSoup [19], an HTML page is parsed to XML. TagSoup does not require the HTML code to be perfectly formed, allowing it to overcome some of the problems many other parsers have with existing ill-formed websites. Second, the package JAXEN [20] provides XPath [21] functionality in order to extract the specific pieces of information. Again the XPath functionality in this package is less susceptible to errors caused by poorly formed XML, when compared to other packages. Third, the named entities in extracted news items are determined using functionality provided by the Stanford Named Entity Recognition and Information Extraction package [22]. Last, once all named entities are determined, all relevant data is stored in an OWL file, which is handled by Jena [23], a Semantic Web framework for Java. The gathered data consists of: the title and body of the news item, the publisher, the date it was published and the named entities, including how many times the named entities were discovered in the article.

V. Evaluation

In this section we evaluate several methods for novelty control. We compare the cosine metric, the Kullback-Leibler divergence, and the Jensen-Shannon divergence, all pairwise and aggregate. As dimensions for news representation we test a vector of only named entities and the vector of all words (excluding stop words).

From the Yahoo! Finance News Archive [24], which offers access to Web articles published by many well-known news providers, we extracted 8,097 news items dated from 28-02-2011 to 08-04-2011. The main stories in our data set are shown in Table I. We left out very small stories (less than 4 items) as these are not interesting for the evaluation due to their lack of informational content. The average number of news items per storyline is approximately 9 and the stories span 2 to 38 days.

We compare all methods to a list of chronologically sorted news items. As evaluation metrics we use Kendall's τ [25] and the Discounted Cumulative Gain [26] to compare a 'golden rank' determined by test users and the rank given by the novelty control methods. Kendall's τ accounts for tied ranks in both lists. The domain of Kendall's τ is from -1 to 1, where 1 indicates that both rankings are the same and

Table I
DESCRIPTION OF MAIN STORIES IN THE NEWS DATABASE

Topic id	Topic description	Number of items
1	Debt crisis Portugal	7
2	Daimler and Rolls-Royce want to jointly buy a supplier	6
3	Oil and gas prices rise by trouble in Middle East	18
4	European Union/European Central Bank debt crisis	20
5	National Football League union meet for Collective Bargaining Agreement negotiations	7
6	Panel to review North Dakota state government pension changes	5
7	Government and big tobacco companies in dispute over proposed ads	8
8	Detroit musicians on strike	6
9	Pennsylvania judge corruption case	5
10	California city officials accused of scam face judge	5

-1 indicates that one ranking is the reverse of the other. The domain of the discounted cumulative gain is from 0 to 1, where 0 indicates a bad correlation between both lists and 1 indicates a perfect correlation. In order to compensate for the bias we would get from tied ranks within the golden rank which are not tied in the rankings provided by the different methods, we modified the existing novelty control methods for Kendall's τ to produce lists with ranks with the same values to the news items as the golden rank, i.e., values 0 (no novelty), 1, 2, and 3 (high novelty) approximately equally spread starting from the most relevant item.

Table II shows an evaluation example, where $Rank_n$ denotes the rank given by novelty control mechanism. $Rank_c$ represents the converted rank given by the novelty control mechanism (by means of a normalization procedure), and $Rank_u$ describes the rank given by the user. Concordant and discordant pairs and the number of tied ranks for Kendall's τ are calculated based on $Rank_c$ and $Rank_u$, while the discounted cumulative gain is calculated based on the ordering of items by the novelty control and $Rank_u$. There are 9 concordant and 2 discordant pairs. There are two groups of ties in both rankings which both count 2 tied ranks. This results in a score of 0.62 for Kendall's τ and 0.95 for the discounted cumulative gain.

We created several test cases, where a test case consists of a list of news items which belong to the same story. Several news items from these lists are considered to be previously read. Three test users were instructed to assign rankings between 0 and 3, (approximately) equally spread, to news items based on their novelty to the seed item and the list of previously read items. The 'golden rank' is constructed based on the rounded average of the users' ratings.

Table II
EVALUATION EXAMPLE

Item	$Rank_n$	$Rank_c$	$Rank_u$
1	0.9	3	3
2	0.8	3	2
3	0.7	2	3
4	0.7	2	1
5	0.4	1	0
6	0.2	0	1

We evaluated the Jensen-Shannon divergence, the Kullback-Leibler divergence, and the cosine similarity. The first two methods are used in combination with word appearance, while we based the cosine metric on TF-IDF weights. Kullback-Leibler divergence is smoothed with linear interpolation smoothing using $\lambda = 0.9$ as suggested by [27]. All methods are used both pairwise and aggregate. As vector dimensions we employed both the vector of all words (stop words excluded) and a vector of only named entities. Tables III and IV show the results. Methods are labeled using the following convention: the first part indicates the divergence distance measure (JS, KL, or COS), the second part indicates if the method is used pairwise or aggregate (P or A) and the last part indicates which dimensions are used for the vector representation (WORDS or NE). For example, pairwise JS divergence with all words as vector is labeled 'JS_P_WORDS'. The chronologically sorted list is labeled 'ORG'. All possible combinations add up to a total of 13 methods.

We discuss the most interesting results based on the data from Tables III and IV. At first sight we can not really determine if the methods work better pairwise or aggregate. The first noteworthy observation is that all methods are outperforming the chronological sorting method for both metrics. Secondly, it can be noted that pairwise and aggregate comparison provide little to no difference in the results. Finally, we can determine the best overall method, which is KL_P_NE for Kendall's τ, and KL_A_NE for the Discounted Cumulative Gain. Other methods which perform well on both metrics are JS_P_NE and JS_P_WORDS.

We can draw more general conclusions based on the averages presented in Tables V and VI. First, we can conclude that the pairwise control mechanism offers slightly better results than the aggregate control. This is due to the fact that the aggregate method by merging all previously read document decreases the quality of novelty control measures. Second, we can observe that named entities outperform all words. This is because named entities capture an important part of the information that depicts the story within a news item. All words contain more information, but are also more susceptible to noise. The words used in the article could be

Table III
KENDALL'S τ OF NOVELTY CONTROL

	Story										
Method	1	2	3	4	5	6	7	8	9	10	AVG
KL_P_NE	-0.671	0.913	0.564	0.463	0.671	1.000	0.689	0.707	-0.548	-0.333	0.346
JS_P_NE	-0.671	0.913	0.665	0.392	0.671	1.000	0.430	0.236	-0.548	-0.333	0.276
COS_A_NE	0.335	0.913	-0.434	0.333	0.894	0.333	0.430	0.236	-0.183	-0.333	0.253
COS_P_NE	0.112	0.913	0.477	0.380	0.112	0.333	0.430	0.236	-0.183	-0.333	0.248
JS_P_WORDS	0.112	0.913	0.564	0.582	0.112	1.000	0.689	0.237	-0.913	-1.000	0.229
KL_A_WORDS	0.671	0.548	-0.362	0.463	0.447	-0.333	0.602	-0.236	-0.913	1.000	0.189
KL_P_WORDS	-0.112	0.913	0.391	-0.071	0.112	1.000	0.689	0.236	-0.913	-0.333	0.169
KL_A_NE	0.000	-0.913	-0.448	0.463	0.224	1.000	0.602	0.236	-0.548	1.000	0.162
COS_A_WORDS	0.335	0.913	-0.231	0.487	-0.112	1.000	0.689	-0.707	-0.548	-0.333	0.149
JS_A_NE	-0.335	0.548	-0.477	0.392	-0.112	1.000	0.689	0.236	-0.548	-0.333	0.106
JS_A_WORDS	0.112	0.913	-0.680	0.392	-0.112	1.000	0.689	-0.236	-0.913	-0.333	0.083
COS_P_WORDS	0.335	0.913	0.391	-0.226	-0.112	0.333	0.689	-0.707	-0.548	-0.333	0.083
ORG	0.335	-0.548	0.391	-0.285	0.112	-1.000	-0.344	0.236	0.183	-1.000	-0.192

Table IV
DISCOUNTED CUMULATIVE GAIN OF NOVELTY CONTROL

	Story										
Method	1	2	3	4	5	6	7	8	9	10	AVG
KL_A_NE	0.817	0.812	0.808	0.874	0.854	0.846	0.993	0.943	0.566	0.734	0.825
KL_A_WORDS	0.885	0.906	0.659	0.738	0.857	0.846	0.993	0.930	0.538	0.768	0.812
JS_P_WORDS	0.689	0.842	0.840	0.843	0.742	0.846	0.823	0.802	0.538	0.768	0.772
JS_P_NE	0.636	0.842	0.848	0.764	0.794	0.846	0.817	0.802	0.566	0.734	0.765
KL_P_NE	0.636	0.842	0.836	0.738	0.794	0.846	0.823	0.814	0.566	0.734	0.763
COS_A_NE	0.779	0.842	0.730	0.779	0.845	0.780	0.806	0.802	0.581	0.673	0.762
KL_P_WORDS	0.756	0.842	0.834	0.672	0.697	0.846	0.823	0.802	0.538	0.768	0.758
COS_P_NE	0.759	0.842	0.831	0.782	0.715	0.780	0.806	0.802	0.581	0.673	0.757
COS_P_WORDS	0.779	0.842	0.827	0.844	0.690	0.822	0.823	0.716	0.566	0.698	0.753
COS_A_WORDS	0.779	0.842	0.721	0.805	0.690	0.846	0.823	0.716	0.566	0.698	0.749
JS_A_NE	0.666	0.832	0.692	0.672	0.697	0.846	0.823	0.802	0.566	0.734	0.743
JS_A_WORDS	0.683	0.842	0.576	0.643	0.690	0.846	0.823	0.780	0.538	0.768	0.731
ORG	0.775	0.728	0.800	0.630	0.742	0.668	0.689	0.802	0.741	0.631	0.721

dependent on the author's writing style, for example. Finally, we can conclude that KL (the only asymmetric distance measure we used) is the best performing method in general.

VI. CONCLUSION

In this paper, we reviewed work in the field of novelty control within Web news stories. Additionally, we have implemented an extensive selection of different novelty control techniques using several distance methods in combination with different representation dimensions and control mechanisms. An evaluation procedure was set up to test the results obtained from the aforementioned implementation for novelty control.

Table V
MEAN KENDALL'S τ RESULTS

Method	Overall	P	A	WORDS	NE
JS	0.173	0.252	0.095	0.156	0.191
KL	0.216	0.257	0.175	0.179	0.254
COS	0.183	0.165	0.201	0.116	0.250

Table VI
MEAN DISCOUNTED CUMULATIVE GAIN RESULTS

Method	Overall	P	A	WORDS	NE
JS	0.753	0.769	0.737	0.752	0.754
KL	0.789	0.760	0.818	0.785	0.794
COS	0.755	0.755	0.755	0.751	0.760

Our evaluation provides a detailed comparison of different novelty control mechanisms currently not existing in the literature. In general methods perform better with named entities than all words as news vector dimensions. The best performing distance measure on average is Kullback-Leibler divergence (with linear interpolation smoothing).

In the future we would like to investigate story detection techniques that take into account the age difference between news items. Also, we plan to investigate the use of semantic approaches [28] for novelty control. While using a named entity in a domain is a good step in this direction, we can go further by using concepts from a domain ontology representative for the available news items for a vector-based document representation. Semantic similarity measures should be useful for novelty control in this context [29].

ACKNOWLEDGMENT

The authors are partially sponsored by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and the FES COMMIT Infiniti project Information Retrieval for Information Services.

REFERENCES

[1] F. Frasincar, J. Borsje, and F. Hogenboom, *E-Business Applications for Product Development and Competitive Growth: Emerging Technologies*. IGI Global, 2011, ch. Personalizing

News Services Using Semantic Web Technologies, pp. 261–289.

[2] F. Frasincar, J. Borsje, and L. Levering, "A Semantic Web-Based Approach for Building Personalized News Services," *International Journal of E-Business Research*, vol. 5, no. 3, pp. 35–53, 2009.

[3] K. Schouten, P. Ruijgrok, J. Borsje, F. Frasincar, L. Levering, and F. Hogenboom, "A Semantic Web-Based Approach for Personalizing News," in *25th Symposium on Applied Computing (SAC 2010)*. ACM, 2010, pp. 854–861.

[4] E. Gabrilovich, S. Dumais, and E. Horvitz, "Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty," in *13th International Conference on World Wide Web (WWW 2004)*. ACM, 2004, pp. 482–490.

[5] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[6] J. G. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in *21st Annual International Conference on Research and Development in Information Retrieval (SIGIR 1998)*. ACM, 1998, pp. 335–336.

[7] J. Allan, C. Wade, and A. Bolivar, "Retrieval and Novelty Detection at the Sentence Level," in *26th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2003)*. ACM, 2003, pp. 314–321.

[8] D. Hiemstra, "Language Models," in *Encyclopedia of Database Systems*. Springer US, 2009, pp. 1591–1594.

[9] B. Fuglede and F. Topsoe, "Jensen-Shannon Divergence and Hilbert Space Embedding," in *IEEE International Symposium on Information Theory (ISIT 2004)*. IEEE, 2004, pp. 31–31.

[10] F. Jelinek and R. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," in *Workshop on Pattern Recognition in Practice*, E. S. Gelsema and L. N. Kanal, Eds. Elsevier, 1980, pp. 381–397.

[11] L. S. Larkey, J. Allan, M. E. Connell, A. Bolivar, and C. Wade, "UMass at TREC 2002: Cross Language and Novelty Tracks," in *11th Text REtrieval Conference (TREC 2002)*, 2002, From: http://trec.nist.gov/pubs/trec11/t11_proceedings.html.

[12] Y. Zhang, J. Callan, and T. Minka, "Novelty and Redundancy Detection in Adaptive Filtering," in *25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*. ACM, 2002, pp. 81–88.

[13] G. Salton and C. S. Yang, "On the Specification of Term Values in Automatic Indexing," *Journal of Documentation*, vol. 29, no. 4, pp. 351–372, 1973.

[14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[15] G. Kumaran and J. Allan, "Text Classification and Named Entities for New Event Detection," in *27th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2004)*. ACM Press, 2004, pp. 297–304.

[16] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, and P. Stein, "OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004," From: http://www.w3.org/TR/owl-ref, 2004.

[17] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF. W3C Recommendation 15 January 2008," From: http://www.w3.org/TR/rdf-sparql-query/, 2008.

[18] Jena Development Team, "ARQ, a SPARQL Processor for Jena," From: http://jena.sourceforge.net/ARQ, 2012.

[19] J. Cowan, "Tagsoup," From: http://ccil.org/~cowan/XML/tagsoup/, 2012.

[20] Jaxen Team, "Jaxen," From: http://jaxen.codehaus.org/, 2012.

[21] W3C, "XML Path Language (XPath)," From: http://www.w3.org/TR/xpath/, 1999.

[22] The Stanford Natural Language Processing Group, "Named Entity Recognition (NER) and Information Extraction (IE)," From: http://nlp.stanford.edu/ner/index.shtml, 2009.

[23] Jena Development Team, "Jena, A Semantic Web Framework for Java," From: http://jena.sourceforge.net/, 2012.

[24] Yahoo!, "Finance News Archive," From: http://biz.yahoo.com/apf/archive.html, 2012.

[25] E. Yilmaz, J. A. Aslam, and S. Robertson, "A New Rank Correlation Coefficient for Information Retrieval." in *31th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2008)*. ACM, 2008, pp. 587–594.

[26] K. Jrvelin and J. Keklinen, "Cumulated Gain-Based Evaluation of IR Techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[27] R. T. Fernández, "The Effect of Smoothing in Language Models for Novelty Detection," in *Proceedings of the BCS IRSG Symposium: Future Directions in Information Access 2007 (FDIA 2007)*, 2007, pp. 11–16.

[28] F. Getahun, J. Tekli, R. Chbeir, M. Viviani, and K. Ytongnon, "Relating RSS News/Items," in *International Conference on Web Engineering (ICWE 2009)*, ser. Lecture Notes in Computer Science, vol. 5648. Springer, 2009, pp. 442–452.

[29] W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom, "Ontology-Based News Recommendation," in *International Workshop on Business intelligencE and the WEB (BEWEB 2010) at Thirteenth International Conference on Extending Database Technology and Thirteenth International Conference on Database Theory (EDBT/ICDT 2010)*, F. Daniel, L. M. L. Delcambre, F. Fotouhi, I. Garrigós, G. Guerrini, J.-N. Mazón, M. Mesiti, S. Müller-Feuerstein, J. Trujillo, T. M. Truta, B. Volz, E. Waller, L. Xiong, and E. Zimányi, Eds. ACM, 2010.