Discovery & Refinement of Scientific Information via a Recommender System

Robert M. Patton, Thomas E. Potok, and Brian A. Worley
Computational Data Analytics
Oak Ridge National Laboratory
Oak Ridge, TN
{pattonrm, potokte, worleyba} @ ornl.gov

***Abstract*—The ability to maintain awareness within a field of research has been a hallmark of scientific expertise for centuries. As diverse scientific information becomes available through various Internet sources, not merely conference proceedings and journals, maintaining scientific awareness becomes a significant challenge. One challenge is how to discover sources that may be of interest. A second challenge lies in finding significant information over many sources. Scanning through hundreds of posts, feeds, and articles a day is very time consuming and error prone. Our approach uses a set of author published papers as seed documents to recommend documents of interest across various Internet sources. This enables 1) discovery of new sources that may be of interest, and 2) refine the information within a source to only the most relevant.**

Keywords-recommender system; text analysis; rss feeds

I. INTRODUCTION

Big data demands the need for intelligent, recommender agents that can enhance a person's situational or domain awareness of their environment. The ability to have a keen awareness and availability of relevant information provides a critical competitive edge. Unfortunately, there is simply too much data streaming too quickly for a person to manually process, analyze, and take action within a reasonable amount of time. This challenge is true in research and academia, as well as industry and government, and has remained a challenge for quite some time [15]. In an attempt to alleviate this challenge, many people subscribe to relevant Internet information. There may be forms of subscriptions with the most common being Really Simple Syndication (RSS), blogs, even Facebook and Twitter. The concept is simple, when new information is posted to the site; a subscriber sees a list of this new information. The subscriber then has the option of following a link to read more. For researchers, the areas to monitor are fairly specific, for example, new research, publication opportunities (e.g., conferences and journals), funding opportunities, the activities of key people in your field, and inspiration for new ideas.

Traditionally, reading key journals and presenting at key conferences and workshops could accomplish this. Now, much of the data has moved to the Internet. The subscriber model is a very useful and successful model for monitoring this data, but it does have some significant drawbacks. In practice, the feeds of new information become quite lengthy, and contain more information than can be practically read. Furthermore, there can be a significant number of items that have little interest to the subscriber. A particular researcher may be strongly interested in another researcher's technical postings, but not interested that researcher's vacation or political postings. Another challenge is how to select the sources to subscribe to. A common model is to subscribe to what your community subscribes to, or to subscribe to the most followed sources. In doing so, it is very difficult to discover a new and interesting source that is not known by another researcher, or known in general. Thus, the ability to find new and relevant information proves critical.

We propose a content-based recommender system was designed and developed called Distribute The Highest Selected Textual Recommendation (DTHSTR) that addresses both of these problems, and is initially developed as a support tool for researchers. However, the flexibility of input allows the system to be adaptable to industry and government use cases as well. Recommender systems enable the filtering of information based on the relevance to or interests of the user. They are often used for e-commerce or entertainment, but rarely, if at all, for the research community in a widespread manner. Our approach attempts to fill this gap for the research community. The following sections will discuss related works, the approach, and an example use case using the approach described.

II. RELATED WORKS

There are many successful approaches to recommending research articles to a user. These approaches are typically applied solely against a corpus of research articles, not on broader information such as call for papers, call for proposals, or science news. Collaborative filtering analyzes information from article reviews of other users as the basis for recommending new articles. There are various machine learning methods to recommend articles based on the citations of other users. Text analysis methods have been used to compare the full or partial content of a set of interesting articles to a set of potential interesting articles using methods such as Term Frequency Inverse Document Frequency (TF-IDF), Latent Semantic Analysis (LSA), and Topic Modeling (TM) [1][6][8][12][13][17][30][31].

In this work our , corpus is a very large and dynamic collection of RSS feed documents, not a research article

 ISBN: 978-1-61208-226-4

corpus. Since there are no reviews and limited citations (e.g., blog comments or Facebook Like), we have focused on comparing the full content of a user's articles against the full content of the RSS feeds. Given the large volume of documents, we are focusing on parallel enhancements to TF-IDF that we believe will perform faster and with less memory requirements than LSA or TM.

In [14], a recommendation system is described based on the similarity to user profiles (i.e., collaborative filtering). The system consists of several databases that contain documents produced over time from a research laboratory, and makes use of the TF-IDF [27][28] term weighting scheme. This work focuses on providing long-term and short-term components of representing a user profile. The long-term representation is created from natural language sentences as a vector space model using the TF-IDF term weighting. The short-term component is based on documents recently downloaded by the user. Profiles are then compared for similarity to each other. The primary drawback to this approach is the database of documents as input, as opposed to a richer, more current on-line data set.

In [33], a personalized recommendation system for scientific and technical papers is described. The system automatically summarizes the technical papers and then performs similarity comparisons to a user's query. While the authors do not clearly describe the details of their approach, the system appears to be dependent on keyword queries provided by the user. Furthermore, another weakness of the approach is the use of automated summarization prior to the similarity comparison, which is very likely to significantly impact the similarity comparison.

In [19], a multi-agent system for recommending scientific documents is presented. Various agents perform different tasks in collaboration with each other via a central profiling agent in order to provide a recommendation to the user. This system appears to be completely ubiquitous in that it simply monitors the users actions as a means of input for recommendations. Recommendations are based on either similarity to other users or similarity to documents that the user has either saved or bookmarked as well as other criteria.

In [6], a recommender system is developed called Scienstein. Like previous works, the system is intended specifically for searching academic papers, and uses both content-based and collaborative-based techniques. In addition, Scienstein performs citation, author, and source analysis as part of its recommendation. Despite the advanced analysis capabilities, there are a few drawbacks. The system deploys are user interface as a desktop application, although a newer version appears to be web-based. In addition, the system is oriented toward academic papers only and does not appear to be expandable to other sources of data.

In [34], an ontology based recommendation system for scholars is described. Unlike the works previously mentioned, this system makes use of search engines such as Google or Yahoo to perform domain specific searches. Results are then processed to extract information using an ontology database, and further refined with a information recommender. Unfortunately, the major drawback to this system is the use of a domain specific ontology. In their work, artificial intelligence ontology was used. In order to apply their system to another domain requires changing the ontology, which may be significantly challenging, or non-existent. In addition, the use of search engines enables the exploration of potentially unknown Internet sources. However, it does not guarantee that information will be monitored consistently from a particular source, and only as regularly as the search engine crawls the source of interest.

In [16], a content-based recommender system using a genetic algorithm is proposed. In this approach, the genetic algorithm is bootstrapped with 20-30 documents that represent a single category of interest. The documents are represented with a vector space model [25] using TF-IDF [27][28] term weighting scheme and cosine similarity metric. The genetic algorithm then evolves to build a classifier for recommending documents to the user. According to the authors, one of the main drawbacks is the bootstrapping process. Another drawback is that TF-IDF approach caused important terms in the user supplied documents to be lower weighted as a result of the documents being very similar to each other. As will be described later, the work described here avoids this through the use of a different term weighting scheme.

In [10], an ontology based recommendation system is developed for browsing web pages and makes use of long-term and short-term preferences. The ontology is built by analyzing book web pages from Amazon, while long-term and short-term preferences are developed by monitoring the web pages viewed by the user. As with the previous ontology-based system, the major drawback is the use of an ontology, which must be learned or developed depending on the domain. Furthermore, the preferences are based on web browsing alone. While this approach may work well for searching books on Amazon, it is not easily or accurately adapted to other domains.

Furthermore, other work [1][5][20][24][35] has focused on developing techniques for filtering RSS feeds. These works focus on RSS feeds in general with a tendency toward news feeds only. Our work uses a manual categorization of the feeds in order to identify content as being associated with a particular aspect of a researcher's workflow (e.g., funding, publications, patents). In addition, they predominantly rely on using some form of supervised learning that requires a training set. Of these works, the work of [35] is the most similar to this work. In [35], a vector space model and TF-IDF term weighting are used while a Rocchio feedback algorithm is used to adapt to the user. Our work uses a different term weighting scheme, and currently, does not support a feedback algorithm.

III. APPROACH

A. Ingest

The DTHSTR system requires two primary sources of input. The first source of input is a list of RSS feeds to be monitored. Currently, the system monitors more than 9,500 feeds from 130 sites. Table I shows a sample of the different sites that are monitored.

 ISBN: 978-1-61208-226-4

TABLE 1. SAMPLE RSS FEEDS

Site	Category
Sciencedaily.com	News
Freshpatents.com	Inventions
Freepatentsonline.com	Inventions
Acm.org	Publications
Ieee.org	Publications
Nejm.org	Publications
Grants.gov	Call for Proposals
Wikicfp.com	Call for Papers

Documents from each feed are monitored and collected on a regular basis. For this particular use case, the second source of input is a set of recent publications for each researcher as a means of representing their research interests. Other documents could be used for different use cases. For this particular work, each publication supplied by the researcher is used individually to provide a reference point for Internet content. This allows the researcher to know which individual publication is most similar to the Internet content.

B. Analysis

In order to process and analyze the input feeds and publications, each document is converted into a collection of terms and associated weights using the vector space model method. The vector space model (VSM) is a recognized approach to document content representation [25] in which the text in a document is characterized as a collection (vector) of unique terms/phrases and their corresponding normalized significance.

Developing a VSM is a multi-step process. The first step in the VSM process is to create a list of unique terms and phrases. This involves parsing the text and analyzing each term/phrase individually for uniqueness using a term weighting scheme. The weight associated with each unique term/phrase is the degree of significance that the term or phrase has, relative to the other terms/phrases. For example, if the term "plan" is common across all or most documents, it will have a low significance, or weight value. Conversely, if "strategic" is a fairly unique term across the set of documents, it will have a higher weight value. The VSM for any document is the combination of the unique term/phrase and its associated weight as defined by a term weighting scheme.

In our approach, the term frequency-inverse corpus frequency (TF-ICF) developed in [22] is used as the term weighting scheme. Over the last three decades, numerous term weighting schemes have been proposed and compared [9][11][26][27]. The primary advantage of using TF-ICF is the ability to process documents in $O(N)$ time rather than $O(N^2)$ like many term weighting schemes, while also maintaining a high level of accuracy. For convenience, the TF-ICF equation is provided here:

$$w_{ij} = \log(f_{ij}) \times \log(N / n_j) \quad (1)$$

In this equation, f_{ij} represents the frequency of occurrence of a term j in document i. The variable N represents the total number of documents in the static corpus of documents, and n_j represents the number of documents in which term j occurs in that static corpus. For a given frequency f_{ij}, the weight, w_{ij}, increases as the value of n decreases, and vice versa. Terms with a very high weight will have a high frequency f_{ij}, and a low value of n.

For the prototype system described here, a corpus of 258,231 documents from a TREC data collection [29] was used for the ICF table. In the ICF table, we store N, which is the total number of documents in the corpus. Also, for each unique term j, after removing the stop words and applying Porter's Stemming Algorithm [21], we store n_j, which is the number of documents in the corpus where term j occurred one or more times. As a result, the task of generating a weighted document vector for a document in a dynamic data stream is as simple as one table lookup. The computational complexity of processing N documents is therefore, $O(N)$.

By using the TF-ICF term weighting scheme, the system avoids the problems with TF-IDF as described in [16]. The ICF component of a user's profile documents (i.e., publications supplied to the system) is compared to a static corpus of news documents rather to each other. This provides a critical benefit: domain specific terms are weighted higher not lower as the TF-IDF scheme would do. This causes the system to be sensitive to different domains, thus enabling its flexibility and use for various domains.

Once a vector representation is created for each document, similarity comparisons can be made. In our approach, a cosine similarity is used to compare two vectors A and B, as shown in (2).

$$\text{Similarity} = (A \cdot B) / (||A|| \, ||B||) \quad (2)$$

Similarity values ranges between 0 and 1, inclusive. A value of 1 means that vectors A and B are identical; while a value of 0 means that they are not alike at all. Recommendations by the DTHSTR system are the documents from each feed that have the highest similarity to the researcher's publications. The number of documents from the feeds can be adjusted either with a threshold setting for the similarity values, or specifying a fixed number of the most similar documents.

C. Output

The DTHSTR system provides the recommendation results via RSS feeds according to categories such as: Call for Proposals, Call for Papers, Inventions, and News. It will also provide a feed of all the results combined into one feed. By providing the results as an RSS feed, this enables the use of any RSS reader program or web component (e.g., as a web part in Microsoft SharePoint) to be used. In addition, this enables the output RSS feed to be used as ingest to another system for further analysis. Finally, this also supports mobile devices.

IV. USE CASE

As an example use case, the authors used one of their own publications entitled "Discovering Potential Precursors of Mammography Abnormalities based on Textual Features, Frequencies, and Sequences" [18]. This publication

 ISBN: 978-1-61208-226-4

discusses research related to temporal analysis of mammograms using Haar wavelets. In [18], the Haar wavelet was used for pattern recognition of precursors to breast cancer or other anomalies. This single publication was used an input to the DTHSTR system running on a single desktop system with two 4-core processors. Example results are shown in Tables II through VI. Table II shows a closely related call for funding proposals from the Air Force Medical Support Agency with a total program funding level of nearly $50 million USD. Table III shows a call for papers for a workshop on data mining healthcare management where the topics include pattern recognition in medical images and data, clinical data analysis, and medical diagnosis. Table IV shows a patent that describes "a method for characterizing signal-vector data using automatic feature selection techniques on wavelet-transformed data to enhance the use of pattern recognition techniques for classification purposes". Table V shows a recommendation for a breast cancer related article from the New England Journal of Medicine. Table VI shows a recommendation for an Association for Computing Machinery news article discussing how machine learning can be used to improve patient diagnosis. As can be seen, with little to no extra effort in their workflow, the authors are now aware of news, patents, funding and publication opportunities that are directly related to their work. The feeds are monitored automatically and the results immediately pushed to the researchers. The researchers no longer have to manually go to each site and perform keyword searches or check each site's feed individually.

TABLE II. CALL FOR PROPOSALS RECOMMENDATION EXAMPLE

Title	Air Force Medical Support Agency (AFMSA/SG8) Modernization Directorate Research / Development and Innovations
Common Terms	Patients, detection, research, diagnosis, performance, medical

TABLE III. CALL FOR PAPERS RECOMMENDATION EXAMPLE 1

Title	DMHM 2012: Third Workshop on Data Mining for Healthcare Management
Common terms	Patients, data, patterns, workshop, detection, diagnosis, analysis

TABLE IV. PATENT RECOMMENDATION EXAMPLE

Title	Method and system for analyzing signal-vector data for pattern recognition from first order sensors
Common terms	Wavelets, coefficient, pre-cursors, haar, patterns, detection, temporal, sampling

TABLE V. JOURNAL ARTICLE RECOMMENDATION EXAMPLE

Title	Breast-cancer Adjuvant Therapy with Zoledronic Acid
Common terms	Patients, breast cancer, lymph, abnormality, diagnosis, analysis

TABLE VI. NEWS ARTICLE RECOMMENDATION EXAMPLE

Title	Better Medicine Through Machine Learning
Common terms	Patients, radiologist, abnormality, diagnostic, patterns, radiology

V. FUTURE WORK

Even with the success of the initial prototype system, there are still areas for improvement. One area is to provide a means for researchers to give feedback to the system on the recommendations. One approach to implementing this is with a semi-supervised approach [3] that relies on the graph Laplacian from spectral graph theory [4]. In this case, a graph is constructed that joins together documents that are similar to each other. The graph can be constructed using a nearest neighbor or similar approach based on proximity in the original feature space. This form of learning would require significantly fewer examples to learn, thus reducing the level of effort by the users to train the system.

Another area for future work involves the automation of finding RSS feeds or other sources of information. The authors manually collected and identified over 130 Internet sources (i.e., sites). Once a site was identified, there was some automation to extract the RSS links in order to eventually collect over 9,500 RSS feeds. Unfortunately, this process was tedious and time consuming. A better approach would be to leverage commercial search engines such as RSS Search Hub [23] to automatically search and collect RSS feeds that are producing information that may be of interest.

VI. SUMMARY

The Internet contains an enormous amount of data that streams faster than can be humanly processed and analyzed. In order for researchers to leverage this data, a recommender system was designed and developed called Distribute The Highest Selected Textual Recommendation (DTHSTR). This system helps fill a critical gap that exists in current technology that can enhance a researcher's awareness in their respective field. The system uses a researcher's recent publications to identify relevant information across more than 9,500 RSS feeds from 130 sources. Documents in the system are represented with a vector space model using the term frequency / inverse corpus frequency (TF-ICF) term weighting scheme. A cosine similarity is used to compare a researcher's publications with those document retrieved from the RSS feeds. Most similar documents are presented to the researcher via an RSS feed. The system is currently deployed and used at Oak Ridge National Laboratory and has been shown to be flexible to various domains as well as enable researchers to quickly maintain awareness of relevant information to their work.

ACKNOWLEDGMENT

Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285; managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR2225. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the

 ISBN: 978-1-61208-226-4

published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

[1] Burkepile, A. and Fizzano, P., "Classifying RSS Feeds with an Artificial Immune System," Second International Conference on Information, Process, and Knowledge Management, pp. 43-47 (2010)

[2] Buckley, C., Singhal, A., and Mitra, M., New retrieval approaches using SMART. In Proc. of the 4th Text Retrieval conference (TREC-4), Gaithersburg (1996)

[3] Chapelle, O., Scholkopf, B., and Zien, A. eds., Semi-Supervised Learning, MIT Press: Cambridge, MA (2006)

[4] Chung, F.R.K., Spectral Graph Theory. American Mathematical Society, Providence, RI (1997)

[5] Garcia, I., and Ng, Y-K., "Eliminating Redundant and Less-Informative RSS News Articles Based on Word Similarity and a Fuzzy Equivalence Relation," 18th IEEE International Conference on Tools with Artificial Intelligence, pp.465-473 (2006)

[6] Gipp, B., Beel, J., and Hentschel, C., "Scienstein: A Research Paper Recommender System," In Proceedings of the International Conferenc on Emerging Trends in Computing, pp. 309-315, (2009)

[7] Hung Chim, and Xiaotie Deng, "Efficient Phrase-Based Document Similarity for Clustering," IEEE Transactions on Knowledge and Data Engineering, vol.20, no.9, pp. 1217-1229, (2008)

[8] Iwasaki, W., Yamamoto, Y., & Takagi, T. (2010). TogoDoc Server/Client System: Smart Recommendation and Efficient Management of Life Science Literature.

[9] Jones, K.S. and Willett, P., Readings in Information Retrieval, Chap. 3. Morgan Kaufmann Publishers, San Francisco, CA, pp. 305-312 (1997)

[10] Kang, J., and Choi, J., "An Ontology-Based Recommendation System Using Long-Term and Short-Term Preferences," 2011 International Conference on Information Science and Applications (ICISA), pp. 1-8, (2011)

[11] Lan, M., Sung, S-Y., Low, H-B, and Tan, C-L., "A comparative study on term weighting schemes for text categorization," In Proc. of the 2005 IEEE International Joint Conference on Neural Networks, vol.1, no., pp. 546- 551, (2005)

[12] Leong, S. (2009). A survey of recommender systems for scientific papers.

[13] McNee, S. M., Kapoor, N., & Konstan, J. A. (2006). Don't look stupid: avoiding pitfalls when recommending research papers.

[14] Nakagawa, A., and Ito, T., "An implementation of a knowledge recommendation system based on similarity among users' profiles," Proceedings of the 41st SICE Annual Conference, pp. 326- 327, (2002)

[15] Pagonis, J., and Sinclair, M., "Evolving personal agent environments to reduce internet information overload: initial considerations," Proceedings of the IEE Colloquium on Lost in the Web: Navigation on the Internet, (1999)

[16] Pagonis, J., and Clark, A.F., "Engene: A genetic algorithm classifier for content-based recommender systems that does not require continuous user feedback," 2010 UK Workshop on Computational Intelligence (UKCI), pp. 1-6, (2010)

[17] Parra, D. (2009). RecULike: Recommending Scientific Articles on CiteULike using variations of Collaborative Filtering Algorithms.

[18] Patton, R.M., and Potok, T. E., "Discovering Potential Precursors of Mammography Abnormalities based on Textual Features, Frequencies, and Sequences", 10th International Conference on Artificial Intelligence and Soft Computing, (2010)

[19] Popa, H.-E.; Negru, V.; Pop, D.; Muscalagiu, I., "DL-AgentRecom - A Multi-Agent Based Recommendation System for Scientific Documents," 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 320-324, (2008)

[20] Pinheiro, W.A., de S. Rodrigues, T., da Silva, M.A.R., da Silva, M.A.N., Silva, M.C.O,; Xexeo, G., and de Souza, J.M., "Autonomic RSS: Discarding Irrelevant News," Fifth International Conference on Autonomic and Autonomous Systems, pp.148-153 (2009)

[21] Porter, M.F., An algorithm for suffix stripping. Program, 14(3), pp. 130-137 (1980)

[22] Reed, J.W., Jiao, Y., Potok, T.E., Klump, B.A., Elmore, M.T., and Hurson, A.R., "TF-ICF: A new term weighting scheme for clustering dynamic data streams," In Proc. of the 5th International Conference on Machine Learning and Applications, pp. 258-263 (2006)

[23] RSS Search Hub, current January 2012, http://www.rsssearchhub.com/

[24] Saha, S., Sajjanhar, A., Gao, S., Dew, R., and Zhao, Y., "Delivering Categorized News Items Using RSS Feeds and Web Services," IEEE 10th International Conference on Computer and Information Technology, pp.698-702 (2010)

[25] Salton, G., Wong, A., and Yang, C.S., "A Vector Space Model for Automatic Indexing," Communications of the ACM, 18(11), pp. 613–620, (1975)

[26] Salton, G., and McGill, M.J., Introduction to Modern Information Retrieval, McGraw Hill Book Co., New York, (1983)

[27] Salton, G. and Buckley, C., Term-weighting approaches in automatic text retrieval. Journal of Information Processing and management, 24(5), pp. 513-523, (1988)

[28] Spark-Jones, K., "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, 28(5), pp. 111-121, (1972)

[29] Text Retrieval Conference data, current January 2012, http://trec.nist.gov/data.html

[30] Torres, R., McNee, S. M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with TechLens.

[31] Wang, C., & Blei, D. M. (2011). Collaborative Topic Modeling for Recommending Scientific Articles.

[32] Xu, H., and Li, C., "A Novel Term Weighting Scheme for Automated Text Categorization," Seventh International Conference on Intelligent Systems Design and Applications, (2007)

[33] Yang, Q., Zhang, S., and Feng, B., "Research on Personalized Recommendation System of Scientific and Technological Periodical Based on Automatic Summarization," First IEEE International Symposium on Information Technologies and Applications in Education (2007)

[34] Yang, S-Y, and Hsu, C-L., "A New Ontology-Supported and Hybrid Recommending Information System for Scholars," 13th International Conference on Network-Based Information Systems, pp. 379-384, (2010)

[35] Zeng, L., Zhang, Y., and Qiu, R. G., "Adaptive User Profiling in Enhancing RSS-based Information Services," IEEE International Conference on Service Operations and Logistics, and Informatics, pp.1-5 (2007)

 ISBN: 978-1-61208-226-4