

# Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang  
*Montclair State University, USA*

Volume III  
K–Pri

Information Science  
**REFERENCE**

**INFORMATION SCIENCE REFERENCE**

Hershey • New York

Director of Editorial Content: Kristin Klinger  
Director of Production: Jennifer Neidig  
Managing Editor: Jamie Snavelly  
Assistant Managing Editor: Carole Coulson  
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavelly, Sean Woznicki  
Cover Design: Lisa Tosheff  
Printed at: Yurchak Printing Inc.

Published in the United States of America by  
Information Science Reference (an imprint of IGI Global)  
701 E. Chocolate Avenue, Suite 200  
Hershey PA 17033  
Tel: 717-533-8845  
Fax: 717-533-8661  
E-mail: [cust@igi-global.com](mailto:cust@igi-global.com)  
Web site: <http://www.igi-global.com/reference>

and in the United Kingdom by  
Information Science Reference (an imprint of IGI Global)  
3 Henrietta Street  
Covent Garden  
London WC2E 8LU  
Tel: 44 20 7240 0856  
Fax: 44 20 7379 0609  
Web site: <http://www.eurospanbookstore.com>

Copyright © 2009 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher.

Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

#### Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.  
p. cm.

Includes bibliographical references and index.

Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.

ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)

1. Data mining. 2. Data warehousing. I. Wang, John,

QA76.9.D37E52 2008

005.74--dc22

2008030801

#### British Cataloguing in Publication Data

A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

*If a library purchased a print copy of this publication, please go to <http://www.igi-global.com/agreement> for information on activating the library's complimentary electronic access to this publication.*

# Measuring the Interestingness of News Articles

**Raymond K. Pon**

*University of California - Los Angeles, USA*

**Alfonso F. Cardenas**

*University of California - Los Angeles, USA*

**David J. Buttler**

*Lawrence Livermore National Laboratory, USA*

## INTRODUCTION

An explosive growth of online news has taken place. Users are inundated with thousands of news articles, only some of which are interesting. A system to filter out uninteresting articles would aid users that need to read and analyze many articles daily, such as financial analysts and government officials.

The most obvious approach for reducing the amount of information overload is to learn keywords of interest for a user (Carreira et al., 2004). Although filtering articles based on keywords removes many irrelevant articles, there are still many uninteresting articles that are highly relevant to keyword searches. A relevant article may not be interesting for various reasons, such as the article's age or if it discusses an event that the user has already read about in other articles.

Although it has been shown that collaborative filtering can aid in personalized recommendation systems (Wang et al., 2006), a large number of users is needed. In a limited user environment, such as a small group of analysts monitoring news events, collaborative filtering would be ineffective.

The definition of what makes an article interesting – or its “interestingness” – varies from user to user and is continually evolving, calling for adaptable user personalization. Furthermore, due to the nature of news, most articles are uninteresting since many are similar or report events outside the scope of an individual's concerns. There has been much work in news recommendation systems, but none have yet addressed the question of what makes an article interesting.

## BACKGROUND

Working in a limited user environment, the only available information is the article's content and its metadata, disallowing the use of collaborative filtering for article recommendation. Some systems perform clustering or classification based on the article's content, computing such values as TF-IDF weights for tokens (Radev et al., 2003). Corso (2005) ranks articles and new sources based on several properties, such as mutual reinforcement and freshness, in an online method. However, Corso does not address the problem of personalized news filtering, but rather the identification of interesting articles for the general public. Macskassy and Provost (2001) measure the interestingness of an article as the correlation between the article's content and real-life events that occur after the article's publication. Using these indicators, they can predict future interesting articles. Unfortunately, these indicators are often domain specific and are difficult to collect for the online processing of articles.

The online recommendation of articles is closely related to the adaptive filtering task in TREC (Text Retrieval Conference), which is the online identification of articles that are most relevant to a set of topics. The task is different from identifying interesting articles for a user because an article that is relevant to a topic may not necessarily be interesting. However, relevancy to a set of topics of interest is often correlated to interestingness. The report by Robertson and Soboroff (2002) summarizes the results of the last run of the TREC filtering task. Methods explored in TREC11 include a Rocchio variant, a second-order perceptron, a SVM, a Winnow classifier, language modelling, probabilistic models of terms and relevancy, and the Okapi Basic Search System.

The recommendation of articles is a complex document classification problem. However, most classification methods have been used to bin documents into topics, which is a different problem from binning documents by their interestingness. Traditional classification has focused on whether or not an article is relevant to a topic of interest, such as the work done in TREC. Typical methods have included the Rocchio (1971) algorithm, language models (Peng et al., 2003), and latent Dirichlet allocation (Newman et al., 2006; Steyvers, 2006). Despite the research done in topic relevancy classification, it is insufficient for addressing the problem of interestingness. There are many reasons why an article is interesting besides being relevant to topics of interests. For example, an article that discusses content that a user has never seen may be interesting but would be undetectable using traditional IR techniques. For example, the events of the September 11 attacks had never been seen before but were clearly interesting. Furthermore, redundant yet relevant articles would not be interesting as they do not provide the user any new information. However, traditional IR techniques are still useful as a first step towards identifying interesting articles.

**MAIN FOCUS**

The problem of recommending articles to a specific user can be addressed by answering what makes an article interesting to the user. A possible classification pipeline is envisioned in Figure 1. Articles are processed in a streaming fashion, like the document processing done in the adaptive filter task in TREC.

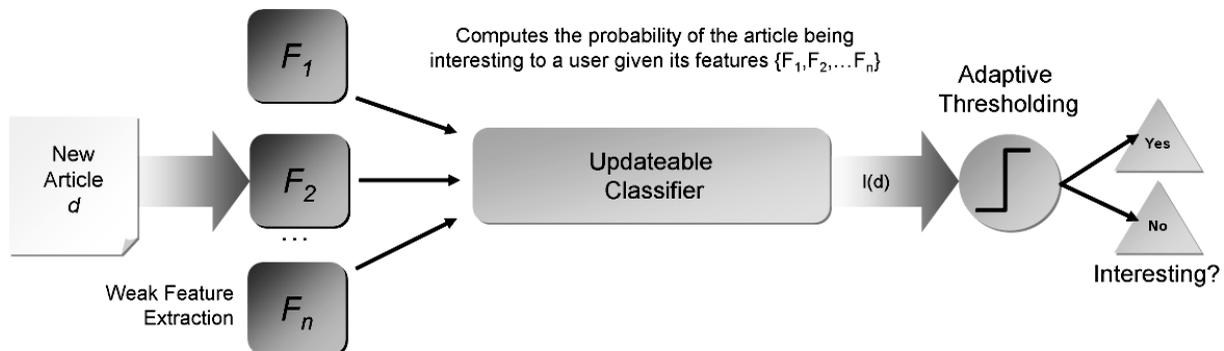
Articles are introduced to the system in chronological order of their publication date. The article classification pipeline consists of four phases. In the first phase, a set of feature extractors generate a set of feature scores for an article. Each feature extractor addresses an aspect of interestingness, such as topic relevancy. Then a classifier generates an overall classification score, which is then thresholded by an adaptive thresholder to generate a binary classification, indicating the interestingness of the article to the user. In the final phase, the user examines the article and provides his own binary classification of interestingness (i.e., label). This feedback is used to update the feature extractors, the classifier, and the thresholder. The process continues similarly for the next document in the pipeline.

**Interestingness Issues**

The “interestingness” of an article varies from user to user and is often complex and difficult to measure. Consequently, several issues arise:

1. There are a variety of reasons why an article is interesting. There is no single attribute of a document that definitively identifies interesting articles. As a result, using only traditional IR techniques for document classification is not sufficient (Pon et al, 2007).
2. Some interestingness features are often contradictory. For example, an interesting article should be relevant to a user’s known interests but should yield new information. On the other hand, random events may be new and unique but may not necessarily be of interest to all users.

Figure 1. Article classification pipeline



3. The breaking news of an important event is difficult to discriminate from the breaking news of an unimportant one.
  4. Because what makes an article interesting varies from user to user, the ideal set of features for a user can not be determined until the system is in use by a user. Useless features will be present in the classification process, which will degrade the performance of a classifier (Forman, 2004), especially its accuracy with classifying on early articles.
  5. The definition of the interestingness may change for a user over time. Consequently, an online learner must be able to adapt to the changing utility of features.
  6. User-feedback must be continually incorporated in the classification process so that any machine learning algorithm can learn efficiently over time what makes an article interesting for a user. A classifier must be incrementally accurate, updateable, and robust against noisy and potentially useless features.
  7. Users are often interested in a multitude of topics that may be drastically different from one another. For example, a user may be interested in news about an election and football. To represent a user using a single profile may not be sufficient while multiple profiles may be costly to maintain (Pon et al., 2007b).
  8. A successful news recommendation system must give accurate recommendations with very little training. Users will deem a system useless if it cannot provide useful recommendations almost immediately.
2. *Uniqueness*: Articles that yield little new information compared to articles already seen may not be interesting. In contrast, an article that first breaks a news event may be interesting. Articles that describe a rare event may also be interesting. For example, Rattigan and Jensen (2005) claim that interesting articles may be produced by rare collaborations among authors. Methods for outlier detection include using mixture models (Eskin, 2000), generating solving sets (Angiulli et al., 2005) and using k-d trees (Chaudhary et al., 2002).
  3. *Source Reputation*: An article's interestingness can be estimated given its source's past history in producing interesting articles. Articles from a source known to produce interesting articles tend to be more interesting than articles from less-reputable sources.
  4. *Writing Style*: Most work using the writing style of articles has mainly been for authorship attribution (Koppel et al., 2006). Instead of author attribution, the same writing style features can be used to infer interestingness. For example, the vocabulary richness (Tweedie & Baayen, 1998) of an article should suit the user's understanding of the topic (e.g., a layman versus an expert). Also writing style features may help with author attribution, which can be used for source reputation, where attribution is unavailable.
  5. *Freshness*: Articles about recent events tend to be labeled as more interesting than articles about older events. Also articles about the same event are published around the time the event has occurred. This may also be the case for interesting events, and consequently interesting articles.
  6. *Subjectivity and Polarity*: The sentiment of an article may also contribute to a user's definition of interestingness. For example, "bad news" may be more interesting than "good news" (i.e., the polarity of the article). Or, subjective articles may be more interesting than objective articles. Polarity identification has been done with a dictionary approach (Mishne, 2005). Others have looked at subjectivity labeling, using various NLP techniques (Wiebe et al., 2004).

### Possible Document Features for Interestingness

There is no single feature that definitively identifies interesting articles. Pon et al. (2007) describes a set of possible aspects regarding interestingness:

1. *Topic Relevancy*: Although an article that is relevant to a topic of interest may not necessarily be interesting, relevancy to such topics is often a prerequisite for interestingness for a certain class of users. Traditional IR techniques can be used for this purpose.

The above list is not an exhaustive list of interestingness features. There is currently ongoing work on

the identification and the measurement of new features that correlate with interestingness.

### Ensembles

Because of the complexity of the problem of recommending articles, a solution to this problem could leverage multiple existing techniques to build a better recommendation system. In other problems, this approach has worked well, such as in webpage duplication (Henzinger, 2006).

One ensemble approach to ranking items, such as articles, is to combine multiple ranking functions through probabilistic latent query analysis (Yan & Hauptmann, 2006). Another approach uses a weighted majority algorithm to aggregate expert advice from an ensemble of classifiers to address concept drift in real-time ranking (Beckier & Arias, 2007). A simpler ensemble approach is taken by Pon et al. (2007a). Different techniques, which are relevant to determining the “interestingness” of an article, are combined together as individual features for a naïve Bayesian classifier. Pon et al. show that this achieves a better “interestingness” judgment. However, naïve Bayesian classifiers assume that features are independent. As discussed earlier, “interestingness” is complex and allows for the possibility of conditionally dependent features. For example, an article may be interesting if it is unique but relevant to topics of interest. The search for an updateable yet efficient and complete classifier for “interestingness” remains open.

Additionally, because the definition of interestingness varies from user to user (Pon et al., 2007a) and may even change over time, it is not possible to use traditional offline feature selection algorithms, such as the ones described by Guyon and Eliseeff (2003), to identify which features are important before deploying the system. So, all features are included for classification. The ideal approach to dealing with this problem is by embedding a feature selection algorithm within an updateable classifier. Some approaches have included using Winnow (Carvalho & Cohen 2006), but lack the generality for handling features with different semantic meanings. Utgoff et al.’s (1997) incremental decision tree algorithm addresses this problem but is not appropriate for an online environment due to its growing storage requirements. A different approach taken by Nurmi and Floreen (2005) identify and remove redundant features using the properties of time

series data. However, this approach is not applicable to articles as articles are not necessarily dependent upon the article that immediately precedes it in the document stream.

### FUTURE TRENDS

With the advent of blogs that specialize in niche news markets, readers can expect to see an explosive growth on the availability of information where only a small fraction may be of interest to them. In contrast to traditional news sources, such as CNN, blogs focus on specific topics that may be of interest to only a handful of users as opposed to the general public. This phenomenon is often referred to as the long tail market phenomenon (Anderson, 2007). Instead of building news filters that cater to the mass public, future research will focus more on personalized news recommendation. Personalization research is also present in other media, as evident in the Netflix Prize competition (2007) and the related KDD Cup 2007 competition (Bennett et al., 2007), in which teams compete to improve the accuracy of movie recommendations.

Traditional corpora, such as the ones used in TREC, are ill equipped to address the problems in personalized news recommendation. Current corpora address the traditional problems of topic relevancy and do not address the problem of interestingness. Furthermore, such corpora are not user-focused. At best, such corpora label articles that a general audience would find to be interesting as opposed to a specific user. Even the Yahoo! news articles used by Pon et al. (2007) address the problem of identifying interesting articles to a large community of users instead of a specific user. Further research in personalized news recommendation will need to be evaluated on a large test data collection that has been collected using many individual users. Such data can be collected by tracking individual user behavior on the Internet or on news bulletin boards, such as Digg (2007).

### CONCLUSION

The online recommendation of interesting articles for a specific user is a complex problem, having to draw from many areas of machine learning, such as feature selection, classification, and anomaly detection. There

is no single technique that will be able to address the problem of interestingness by itself. An ensemble of multiple techniques is one possible solution to addressing this problem. Because of the growth of research in recommendation systems, more user-focused test collections should be made available for system evaluation and comparison.

## REFERENCES

- Anderson, C. (2007). Calculating latent demand in the long tail. *Proceedings of the 13<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press.
- Angiulli, F., Basta, S., & Pizzuti, C. (2005). Detection and prediction of distance-based outliers. *Proceedings of the 2005 ACM Symposium on Applied Computing*, pages 537–542. New York: ACM Press.
- Becker, H. & Arias, M. (2007). Real-time ranking with concept drift using expert advice. *Proceedings of the 13<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 86–94. New York: ACM Press.
- Bennett, J., Elkan, C., Liu, B., Smyth, P., & Tikk, D., editors (2007). *Proceedings of KDD Cup and Workshop 2007*. ACM SIGKDD.
- Carreira, R., Crato, J. M., Goncalves, D., & Jorge, J. A. (2004). Evaluating adaptive user profiles for news classification. *Proceedings of the 9<sup>th</sup> International Conference on Intelligent User Interface*, pages 206–212. New York: ACM Press.
- Carvalho, V. R. & Cohen, W. W. (2006). Singlepass online learning: Performance, voting schemes and online feature selection. *Proceedings of the 12<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 548–553. New York: ACM Press.
- Chaudhary, A., Szalay, A. S., & Moore, A. W. (2002). Very fast outlier detection in large multidimensional data sets. In *DMKD*.
- Corso, G. M. D., Gulli, A., & Romani, F. (2005). Ranking a stream of news. *Proceedings of the 14th International Conference on World Wide Web*, pages 97–106. New York: ACM Press.
- Digg (2007). *Digg*. Retrieved September 21, 2007, from <http://www.digg.com>
- Eskin, E. (2000). Detecting errors within a corpus using anomaly detection. *Proceedings of the 1<sup>st</sup> Conference on North American Chapter of the Association for Computational Linguistics*, pages 148–153. San Francisco: Morgan Kaufmann Publishers Inc.
- Forman, G. (2004). A pitfall and solution in multi-class feature selection for text classification. *Proceedings of the 21<sup>st</sup> International Conference on Machine Learning*, page 38.
- Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- Henzinger, M. (2006). Finding near-duplicate web pages: a large-scale evaluation of algorithms. *Proceedings of the 29<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 284–291. New York: ACM Press.
- Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship attribution with thousands of candidate authors. *Proceedings of the 29<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–660. New York: ACM Press.
- Macskassy, S. A. & Provost, F. (2001). Intelligent information triage. *Proceedings of the 24<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–326. New York: ACM Press.
- Mishne, G. (2005). Experiments with mood classification in blog posts. *Style2005 - the 1<sup>st</sup> Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR 2005*.
- Netflix (2007). *Netflix prize*. Retrieved September 21, 2007, from <http://www.netflixprize.com/>
- Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. *IEEE International Conference on Intelligence and Security Informatics*.
- Nurmi, P. and Floreen, P. (2005). Online feature selection for contextual time series data. *PASCAL Subspace, Latent Structure and Feature Selection Workshop*, Bohinj, Slovenia.

Peng, F., Schuurmans, D., & Wang, S. (2003). Language and task independent text categorization with simple language models. *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 110–117. Morristown, NJ: Association for Computational Linguistics.

Pon, R. K., Cardenas, A. F., Buttler, D., & Critchlow, T. (2007a). iScore: Measuring the interestingness of articles in a limited user environment. *IEEE Symposium on Computational Intelligence and Data Mining 2007*, Honolulu, HI.

Pon, R. K., Cardenas, A. F., Buttler, D., & Critchlow, T. (2007b). Tracking multiple topics for finding interesting articles. *Proceedings of the 13<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–569. New York: ACM Press.

Radev, D., Fan, W., & Zhang, Z. (2001). Webinessence: A personalised web-based multi-document summarisation and recommendation system. *Proceedings of the NAACL-01*, pages 79–88.

Rattigan, M. & Jensen, D. (2005). The case for anomalous link detection. *4th Multi-Relational Data Mining Workshop, 11<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Robertson, S. & Soboroff, I. (2002). The TREC 2002 filtering track report. In *TREC II*.

Rocchio, J. (1971). *Relevance feedback in information retrieval*, pages 313–323. Prentice-Hall.

Steyvers, M. (2006). *Latent semantic analysis: A road to meaning*. Laurence Erlbaum.

Tweedie, F. J. & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.

Utgoff, P. E., Berkman, N. C., & Clouse, J. A. (1997). Decision tree induction based on efficient tree restructuring. *Machine Learning*, 29(1).

Wang, J., de Vries, A. P., & Reinders, M. J. T. (2006). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. *Proceedings of the*

*29<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 501–508. New York: ACM Press.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Yan, R. & Hauptmann, A. G. (2006). Probabilistic latent query analysis for combining multiple retrieval sources. *Proceedings of the 29<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–331. New York: ACM Press.

## KEY TERMS

**Conditionally Dependent Features:** Features whose values are dependent upon the values of other features.

**Ensemble:** The combination of multiple techniques to achieve better results for a common task.

**General Audience Recommendation:** Recommendations made for the mass public, usually related to what is popular.

**Interestingness:** How interesting the referred item is to a specific user. This measure is complex and subjective, varying from user to user.

**Online Feature Selection:** The problem of selecting a subset of useful features from a set of given features for online classification by machine learning algorithms. As instances are classified sequentially, the appropriate set of features is selected for classifying each instance.

**Online News Recommendation:** The problem of recommending news articles to a specific user by machine learning algorithms. Such algorithms must provide a recommendation for an article when it arrives in a document stream in real-time. Once a decision on an article is made, the decision cannot be changed.

**User-Focused Recommendation:** Recommendations made for a specific user or a niche community.