World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:11, 2007

Searching for Similar Informational Articles in the Internet Channel

Sung Ho Ha, Seong Hyeon Joo, and Hyun U. Pae

***Abstract*—**In terms of total online audience, newspapers are the most successful form of online content to date. The online audience for newspapers continues to demand higher-quality services, including personalized news services. News providers should be able to offer suitable users appropriate content. In this paper, a news article recommender system is suggested based on a user's preference when he or she visits an Internet news site and reads the published articles. This system helps raise the user's satisfaction, increase customer loyalty toward the content provider.

***Keywords*—**Content classification, content recommendation, customer profiling, documents clustering.

I. Introduction

IN terms of total online audience, newspapers are the most successful form of online content to date. Online newspapers have offered the opportunity for deeper content, multimedia, and interactivity that is not possible in the traditional print newspapers [1].

Online newspapers are the top choice for local news and information for internet users. The online audience for newspapers continues to grow in both numbers and sophistication, demanding higher-quality online delivery and more services: searchable archives, personalized news services, intermediary services, and audio and video coverage of special or local events [2].

However, most online newspapers still offer all users the same content, failing to satisfy an individual user's needs. News providers should be able to offer suitable users suitable content [3]. To do so, they must be able to identify the customers, predict their interests, determine the appropriate content, and deliver it in a personalized format during customers' online sessions. The recommender system of digital information suggests online content (news articles), based on a user's preference when he or she visits an Internet news site and reads the published articles. This system creates a one-to-one relationship between the content provider and the user, raises the user's satisfaction, and increases loyalty toward the content provider.

Manuscript received October 15, 2007.

S. Ha is with Kyungpook National University, Daegu, 702-701, Korea (phone: +82-53-950-5440; fax: +82-53-950-6247; e-mail: hsh@knu.ac.kr).

S. Joo is with Kyungpook National University, Daegu, 702-701, Korea (e-mail: sideas@daum.net).

H. Pae is with Kyungpook National University, Daegu, 702-701, Korea (e-mail: ape00@nate.com).

II. Literature Review

There have been numerous research efforts on content personalization in Web-based applications. Kienle *et al*. (1997) developed a system named DeNews which implements text and natural language processing techniques to personalize news to the specific interests of a user [4]. Bharat, Kamba, and Albers (1998) presented an interactive, personalized newspaper, which builds up user profiles in order to tailor the newspaper content and layout to each user's preferences [5]. Konstantas and Morin (2000) developed an agent-based framework, Hypermedia Electronic Publishing, for the commercial dissemination of electronic documents [6].

Jokela, Turpeinen, Kurki, Savia, and Sulonen (2001) explained the SmartPush system which provides personalized news to a customer on the Web [7]. Kohrs and Merialdo (2001) applied collaborative filtering for user-adapted websites which the audiences can personalize by customizing content [8]. Kuo and Shan (2002) presented a personalized content-based music filtering system to support music recommendations based on a user's preference for a melody style [9]. Shi, Collins, and Karamcheti (2003) captured the characteristics of dynamic content from several representative news and e-commerce sites to develop a dynamic content emulator [10].

Zhang (2003) proposed a generic framework for delivering personalized and adaptive content to mobile users, which can be applied to context-aware mobile services [11]. Tseng, Lin, and Smith (2004) designed a video personalization and summarization system, incorporating the usage environment to select the optimal set of desired contents according to user preferences [12]. Bezerra and Carvalho (2004) presented a recommender system through which each user profile is modeled by summarizing the information taken from a set of items the user has previously evaluated [13].

Boavida, Cabaco, and Correia (2005) proposed a system, VideoZapper, for delivering personalized video content based on its properties and on the past experience of other users with that material [14]. Liu and Lin (2005) presented an incremental text mining technique to efficiently identify a user's current interest by mining the user's information folders, which hold the user's information of interest [15]. Li, Lu, and Xuefeng (2005) explored a hybrid collaborative filtering method based on item and user, which identifies similar movies to a target movie and determines neighbor users of the active user for the target movie [16].

Wei, Chiang, and Wu (2006) established personalized

document clustering techniques that consider individual preferences to support personalization in document categorization [17]. Li, Myaeng, and Kim (2007) described a collaborative music recommender system based on an item-based probabilistic model, where items are classified into groups and predictions are made for users by considering the user ratings [18]. Kazienko and Adamski (2007) developed the AdROSA system for automatic web banner personalization, which integrates web usage and content mining techniques to enable online and personalized advertising [19]. Hsu (2008) developed an online personalized English-learning recommender system, which provides students with appropriate reading lessons that fit their different interests [20].

III. Internet News Recommender System

In order to find similar digital information on the Internet, we have developed the following system architecture (refer to Fig. 1 below).

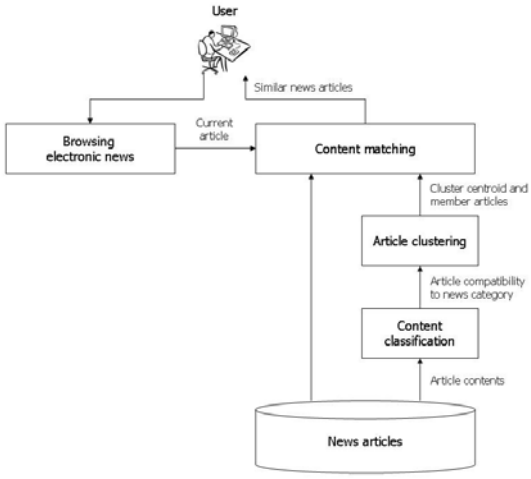

Fig. 1 The architecture of the recommender system

The system performs the Bayesian classification, builds article clusters, and performs content matching that is used in an online recommendation environment.

A. Bayesian Classification of News Contents

Most electronic news sites supply digital news on the Web according to a predetermined set of categories. Some of the news, however, belongs to multiple categories at the same time. To measure each article's degree of relevance to each established category, a text mining technique, the Bayesian classification, classifies digital news according to the vocabulary which occurs in any news article [21-24].

The Bayesian classification is a learning method based on the Bayesian theorem. Let D be a collection of m documents d_i, $i = 1…m$, each one belonging to categories c_j, $j = 1…n$. The Bayesian theorem estimates the probability $p(c_j \mid d_i)$ that a document d_i is in category c_j:

$$p(c_j \mid d_i) = \frac{p(d_i \mid c_j) \cdot p(c_j)}{\sum_{l=1}^{n} p(d_i \mid c_l) p(c_l)} \quad (1)$$

where $p(c_j)$ denotes a prior probability per category, and $p(d_i \mid c_j)$ the likelihood of observing document d_i in category c_j.

The vectors of posterior probabilities, $p(c_j \mid d_i)$, calculated by the Bayesian probabilistic model, reveal that a news article can belong to one or more document categories. These vectors are used to cluster articles.

B. News Clustering

Using the vectors of posterior probabilities calculated by the Bayesian classifier, a Self-Organizing Map (SOM) is chosen to cluster these articles into numerous segments. The SOM tries to uncover patterns in the set of the posterior probabilities by category, and clusters the news articles into distinct segments [25].

The SOM algorithm uses competitive learning. When an input pattern is imposed on the neural network, the algorithm selects the output node with the smallest Euclidean distance between the presented input pattern vector and its weight vector. Only this winning neuron generates an output signal from the output layer. Because learning involves a weight-vector adjustment, only the neurons in the winning neuron's neighborhood can learn with this particular input pattern. They do this by adjusting their weights closer to the input vector as follows (2):

$$w_j(n+1) = \begin{cases} w_j(n) + \eta(n)[x(n) - w_j(n)], \; j \in N(n) \\ w_j(n), \; otherwise \end{cases} \quad (2)$$

where w_j is a weight vector, x is the presented input vector, η is a learning rate, and N is a neighborhood function. As learning proceeds, each news article within a segment tends to have a similar content, and articles in different segments have a different content.

C. News Matching

In order to prepare appropriate digital news based on the user's browsing behavior, the recommender system calculates the Euclidean distance between the current-browsing news and each member article in the chosen segment where the current-browsing article belongs. The system, then, attempts to choose the nearest articles. The nearer the distance is, the more similar the news is to the current-browsing article.

Formally, a distance measure between the current-browsing article and each news article (d_i) is defined as (3):

$$\sqrt{\sum_{j=1}^{n} (PPC_j - PPM_j)^2} \quad (3)$$

where *PPC* represents the current-browsing article expressed by a vector of posterior probabilities in Euclidean n-space, *PPM* denotes each member article expressed by a vector of posterior probabilities, and n represents the number of news categories. The system uses (3) to determine the

recommendable digital news which is forwarded to the customer.

IV. Application of the System

The recommender system, which has been developed in the form of a prototype system, has been applied to an English news site run by a Korean newspaper company. The web site has classified news into six main categories: business, culture, nation, opinion, sports, and technology. The experimental database contained 8,071 sample news articles during seven months from Nov. 2005, which were randomly sampled from a real-world database. 1,481 articles were extracted from the category Business, 936 articles from Culture, 2,345 articles from Nation, 1,290 articles from Opinion, 860 articles from Sports, and 1,159 articles from Technology.

The Bayesian classifier classified the news articles. Table I shows the results of news classification.

TABLE I
Posterior Probabilities Derived by a Bayesian Classifer

Article ID	Posterior probability by category					
	Biz	Cul	Nat	Opi	Spo	Tec
d_{10327}	0.77	0.00	0.00	0.00	0.00	0.23
d_{20359}	0.00	0.89	0.11	0.00	0.00	0.00
d_{50936}	0.00	0.00	0.00	0.00	1.00	0.00

Biz stands for business category, Cul for culture, Nat for nation, Opi for opinion, Spo for sports, and Tec for technology.

Title of article d_{10327}: "Online retailers less bullish about business upturn"; d_{20359}: "Association holds T-shirt design contest"; d_{50936}: "Lee Young-pyo ready to move on."

The recommender system divided news articles into several segments on the basis of classification probabilities (as shown in Table I), and assigned each article to the resulting news segments. Table II summarizes ten article-segments derived from a four-by-four SOM, and the centroid of each segment expressed by vectors of the news categories.

TABLE II
News Segments and Their Centroids

Segment	Centroid of segment					
	Biz	Cul	Nat	Opi	Spo	Tec
C_nat	0.00	0.00	1.00	0.00	0.00	0.00
C_biz	0.99	0.00	0.00	0.00	0.00	0.00
C_opi	0.00	0.00	0.00	1.00	0.00	0.00
C_cul	0.00	1.00	0.00	0.00	0.00	0.00
C_tec	0.00	0.00	0.00	0.00	0.00	1.00
C_nat_biz	0.02	0.00	0.96	0.01	0.00	0.01
C_nat_biz_cul	0.14	0.07	0.68	0.01	0.03	0.07
C_opi_nat	0.01	0.02	0.45	0.52	0.00	0.00
C_biz_tec	0.59	0.00	0.09	0.00	0.00	0.32
C_tec_biz	0.19	0.01	0.05	0.00	0.00	0.75

A current-browsing page contains a user's interests, which can be captured during a current interaction (i.e., browsing a specific Web page) in the Internet news site. The recommendation system provides documents that have a smaller distance to the document that a user is currently browsing (i.e., most similar to keywords of the browsing document).

For example, when a user is browsing an article, such as article ID 20359, which belongs to content segment C_cul_nat and is titled "Association holds T-shirt design contest," the system finds the nearest articles within that segment, C_cul_nat. Table III lists the six nearest articles and their characteristics, such as title, distance within the target content segment, and compatibility expressed by the posterior probabilities.

TABLE III
Top Six Articles Recommendable with the Nearest Distance

Rank	Article ID				Distance	
	Biz	Cul	Nat	Opi	Spo	Tec
1		*20980*			0.0002	
	0.00	0.90	0.10	0.00	0.00	0.00
2		*20520*			0.0003	
	0.00	0.88	0.12	0.00	0.00	0.00
3		*20537*			0.0006	
	0.00	0.87	0.13	0.00	0.00	0.00
4		*20594*			0.0015	
	0.00	0.86	0.14	0.00	0.00	0.00
5		*20409*			0.0016	
	0.01	0.86	0.13	0.00	0.00	0.00
6		*50785*			0.0074	
	0.00	0.83	0.17	0.00	0.00	0.00

Title of article d_{20980}: "Le Book on the second floor"; d_{20520}: "Gugak station aims to make global splash"; d_{20537}: "La Traviata"; d_{20594}: "Memorial features Um Hong-Gil's ascent of all 8,000-m peaks"; d_{20409}: "The forecast is snow"; d_{50785}: "Host of morning show becomes public relations envoy."

V. Conclusion

This study developed a news recommendation system, which provided personalized online content, based on a user's preference. The system performed a Bayesian classification, news clustering, and a content matching function that was used in an online recommendation environment.

Although this study has several characteristics over other research on content classification and clustering, future research can extend this work. First, when classifying content, this study uses a Bayesian classification. Although the algorithm has several advantages, including being simple and easy-to-use, a variety of classification techniques should be deployed and the recommendation performance among them should be compared. Second, further research must include a performance comparison between this recommender system and other ones by measuring user satisfaction and validating user profiles in a real-world situation.

References

[1] C. Ihlstrom and J. Palmer, "Revenues for online newspapers: owner and user perceptions," *Electronic Markets*, vol. 12, no. 4, pp. 228-236, 2002.

[2] R. Shah, R. Jain, and F. Anjum, "Efficient dissemination of personalized information using content-based multicast," *Proc. Of IEEE-Infocom*, 2002, Jun. 23-27.

[3] V. K. Gupta, S. Govindarajan, and T. Johnson, "Overview of content management approaches and strategies," *Electronic Markets*, vol. 11, no. 4, pp. 281-288, 2001.

[4] S. Kienle, S. Lingler, W. Kraas, A. Offenhausser, W. Knol, G. Jung, A. L. K. Wee, C. T. Loong, and J. C. Tiak, "DeNews – a personalized news system," *Expert Systems with Applications*, vol. 13, no. 4, pp. 249-257, 1997.

[5] K. Bharat, T. Kamba, and M. Albers, "Personalized, interactive news on the Web," *Multimedia Systems*, vol. 6, pp. 349-358, 1998.

[6] D. Konstantas and J.-H. Morin, "Agent-based commercial dissemination of electronic information," *Computer Networks*, vol. 32, pp. 753-765, 2000.
[7] S. Jokela, M. Turpeinen, T. Kurki, E. Savia, and R. Sulonen, "The role of structured content in a personalized news service,", *Proc. of the 34th Hawaii International Conference on System Sciences*, 2001, Jan. 3-6, pp. 1-10.
[8] A. Kohrs and B. Merialdo, "Creating user-adapted Websites by the use of collaborative filtering," *Interacting with Computers*, vol. 13, pp. 695-716, 2001.
[9] F.-F. Kuo and M.-K. Shan, "A personalized music filtering system based on melody style classification," *Proc. of the 2002 IEEE International Conference on Data Mining*, 2002, pp. 649-652.
[10] W. Shi, E. Collins, and V. Karamcheti, "Modeling object characteristics of dynamic Web content," *Journal of Parallel and Distributed Computing*, vol. 63, pp. 963-980, 2003.
[11] D. Zhang, "Delivery of personalized and adaptive content to mobile devices: a framework and enabling technology," *Communications of the Association for Information Systems*, vol. 12, pp. 183-202, 2003.
[12] B. L. Tseng, C.-Y. Lin, and J. R. Smith, "Video personalization and summarization system for usage environment," *Journal of Visual Communication & Image Representation*, vol. 15, pp. 370-392, 2004.
[13] B. L. D. Bezerra and F. A. T. Carvalho, "A symbolic approach for content-based information filtering," *Information Processing Letters*, 92, pp. 45-52, 2004.
[14] M. Boavida, S. Cabaco, and N. Correia, "VideoZapper: a system for delivering personalized video content," *Multimedia Tools and Applications*, vol. 25, pp. 345-360, 2005.
[15] R.-L. Liu and W.-J. Lin, "Incremental mining of information interest for personalized web scanning," *Information Systems*, vol. 30, pp. 630-648, 2005.
[16] Y. Li, L. Lu, and L. Xuefeng, "A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in e-commerce," *Expert Systems with Applications*, vol. 28, pp. 67-77, 2005.
[17] C.-P. Wei, R. H. L. Chiang, and C.-C. Wu, "Accommodating individual preferences in the categorization of documents: a personalized clustering approach," *Journal of Management Information Systems*, vol. 23, no. 2, pp. 173-201, 2006.
[18] Q. Li, S. H. Myaeng, and B. M. Kim, "A probabilistic music recommender considering user opinions and audio features," *Information Processing and Management*, vol. 43, pp. 473-487, 2007.
[19] P. Kazienko and M. Adamski, "AdROSA – Adaptive personalization of web advertising," *Information Sciences*, vol. 177, pp. 2269-2295, 2007.
[20] M.-H. Hsu, "A personalized English learning recommender system for ESL students," *Expert Systems with Applications*, vol. 34, pp. 683-688, 2008.
[21] M.-F. Moens, *Automatic indexing and abstracting of document texts*, MA: Kluwer Academic Publishers, 2000.
[22] G. Kowalski and M. T. Maybury, *Information storage and retrieval systems: theory and implementation*, MA: Kluwer Academic Publishers, 2000.
[23] S. M. Weiss, N. Indurkhya, T. Zhang, and F. J. Damerau, *Text mining: predictive methods for analyzing unstructured information*, NY: Springer, 2007.
[24] M. Konchady, *Text mining application programming*, Charles River Media, 2006.
[25] M. Mohammadian, *Intelligent agents for data mining and information retrieval*, PA:Idea Group Publishing, 2004.