Collaborative Filtering to Improve Navigation of Large Radiology Knowledge Resources

Charles E. Kahn, Jr., M.D., M.S.

Objective. Collaborative filtering is a knowledge-discovery technique that can help guide readers to items of potential interest based on the experience of prior users. This study sought to determine the impact of collaborative filtering on navigation of a large, Web-based radiology knowledge resource. **Materials and Methods.** Collaborative filtering was applied to a collection of 1,168 radiology hypertext documents available via the Internet. An item-based collaborative filtering algorithm identified each document's six most closely related documents based on 248,304 page views in an 18-day period. Documents were amended to include links to their related documents, and use was analyzed over the next 5 days. **Results.** The mean number of documents viewed per visit increased from 1.57 to 1.74 ($P < 0.0001$). **Conclusions.** Collaborative filtering can increase a radiology information resource's utilization and can improve its usefulness and ease of navigation. The technique holds promise for improving navigation of large Internet-based radiology knowledge resources.

KEY WORDS: Collaborative filtering, World Wide Web, navigation, teaching files, digital libraries, image libraries, information resources, recommender systems, social filtering, word-of-mouth recommendation, mass customization, knowledge management

INTRODUCTION

Internet-based technologies—particularly the World Wide Web—have succeeded in making available large volumes of images and text on demand. It is now common for radiologists to develop collections of thousands of data items. As a result, users are often faced with the challenge of identifying and navigating to relevant information within a sea of data. Collaborative filtering is an automated approach to help readers find the information that is most valuable by using the experience and preferences of prior users.¹⁻³ This approach, also called “social filtering” and

“recommender systems,” relates the experiences of users to “filter” or select those choices that may be most relevant.⁴ We sought to determine whether collaborative filtering would improve navigation of a radiology information resource, which would be manifested as an increased utilization of the resource.

MATERIALS AND METHODS

CHORUS (<http://chorus.rad.mcw.edu>) is a mature, Web-based radiology resource that describes diseases, imaging findings, and pertinent anatomy, physiology, pathology, and technical information.⁵ It comprises a home page, a set of topical index pages, and a collection of 1,168 brief hypertext documents by 19 authors. Documents are indexed by title and by organ system, and are richly hyperlinked to related documents within the site. CHORUS has a sufficiently large collection of documents and receives enough Web traffic for statistically valid sampling (e.g., 1.4 million page views in March 2004; mean, 44,927 per day; M. Sauer, personal communication). A sample document is shown in Figure 1A. A record of all CHORUS documents accessed between April 1 and April 18, 2003, was obtained from the log file of server activity. Only hits to the notecard documents were included; hits to the home page, topical index pages, and search pages were excluded from the tally. The data were parsed to extract

From the Department of Radiology, Medical College of Wisconsin, 9200 W. Wisconsin Ave., Milwaukee, WI 53226, USA

Submitted to the Journal of Digital Imaging

Correspondence to: Charles E. Kahn, Jr., M.D., M.S., Department of Radiology, Medical College of Wisconsin, 9200 W. Wisconsin Ave., Milwaukee, WI 53226, USA; tel: +1-414-805-2173; fax: +1-414-259-9290; e-mail: kahn@mcw.edu

Copyright © 2004 by SCAR (Society for Computer Applications in Radiology)

Online publication 19 April 2005

doi: 10.1007/s10278-004-1910-9

the client IP address, a timestamp, and the ID of the requested document.

A “visit” was defined as a series of one or more page views from the same client IP address, where no more than 60 minutes were allowed between retrieval of consecutive pages. This time limit was employed to distinguish between distinct uses of the database by one or more users with the same client IP address. Each visit was treated as a set of distinct documents. The order in which documents were viewed was not considered. Because users frequently followed a document’s outbound hypertext links, we sought to capture the natural “clusters” of documents, which might include those items from which a document might have inbound links. Retrieval of any document more than once within a specific visit was ignored. Visits of 2 to 10 documents were included in the analysis. Visits of more than 10 documents were excluded because they were atypical of the usual browsing pattern and frequently represented scans by Web-crawling robots.

Preliminary analysis of the server logs indicated that the majority of visits entailed viewing of only one notecard document. An item-based recommendation approach was selected due to its advantages of simplicity and speed, and to overcome the sparseness of data. For each included visit, an index of all unordered pairs of documents was generated. The index was sorted in descending order by frequency of that distinct pair. For each document, the top six related documents were identified.

Each CHORUS document was altered to include links to its six most closely related documents. Figure 1B shows a

document modified to include recommendations. The modified documents were published to the original Web site. The recommendations remained fixed and were not updated manually or automatically during the trial period or subsequently.

An informal audit of the recommendations was conducted. Forty of the 1,168 modified CHORUS documents were selected at random. The author reviewed the recommendations to assess the quality of the recommendations to assure that they did indeed point to closely related documents.

The Web server log was then monitored for the subsequent 5-day period (May 1 through May 5, 2003), and the number of document page views was recorded. This relatively short period was selected to minimize the possibility that any factors related to Web search engines (such as a change in relative page rankings) would affect the use of the site. The number and size of visits recorded during this period were analyzed. The modified Web pages (with recommendations) remained on the Web site after the trial period.

To confirm the impact of recommendations on system use, a second “reverse” trial was conducted 1 year later. Web site activity from May 1 to May 31, 2004 (with recommendations) was recorded. On June 1, 2004, the published documents were returned to their original appearance (without recommendations); Web site activity was recorded from June 1 to June 30, 2004. On July 1, 2004, the modified documents (with recommendations) were restored to the Web site. The visit size was compared between May 2004 (with recommendations) and June 2004 (without recommendations).

Fig 1. (A) Example CHORUS document describes criteria for and causes of diffuse thickening of the gallbladder wall. The document contains links to the topical index (“Liver and biliary system”) and to other CHORUS documents. (B) CHORUS document with recommendations. The “Related documents” are hyperlinks to the six most closely related items in the collection. Note that although the document has direct “outbound” hyperlinks to the first three of these items, the last three do not have direct links, but are topically related. This example shows how collaborative filtering can identify related documents that may not be linked directly to the document at hand.

Fig 2. Histogram shows the distribution of visit sizes in the control period (without recommendations, used for training) and the intervention period (with recommendations). There is a shift toward larger visit sizes.

RESULTS

There were 248,304 notecard page views during the April 1–18, 2003, control period (mean, 13,795 page views per day), which represented a total of 107,890 user visits. The majority of visits (68.7%) involved retrieval of only one document. Visits in which more than 10 distinct documents were retrieved accounted for 1.0% of all visits. Visits of 2 to 10 documents (32,706 visits; 30.3%) were included in the analysis.

During the 5-day intervention period, there were 67,710 document page views (mean, 13,542 page views per day). We explored whether the presence of the recommendations for related documents increased the number of pages viewed. As shown in Figure 2 below, there was a statistically significant shift to larger visit sizes during the intervention period; a chi-square test yielded $p \ll 10^{-10}$. The mean visit size (among visits of 1 to 10 documents only) increased from 1.57 to 1.74 ($P < 0.0001$).

In the 40 documents audited, the recommendations did indeed bear close relation to the index document and were considered appropriate. A sample of the recommendations is shown in Table 1. As expected, many of the documents had hyperlinks from the target document. In addition, the filtering process identified many “backward links,” that is, documents that pointed back to the document under consideration.

The “reverse” experiment conducted in May–June 2004 confirmed the results of the initial trial. The mean number of pages viewed per visit was 1.67 in May 2004, when the published documents contained the links to related documents, compared to 1.53 in June 2004, when the links were removed. Thus, utilization of the Web site was greater when links to related documents were offered ($P < 0.0001$).

Table 1. Recommendations for the first six documents

00001 Acanthosis nigricans	00004 Acro-osteolysis
00003 Achalasia	00334 Progeria (Hutchinson–Gilford) syndrome
00857 Barrett esophagus	00831 Scleroderma
00071 Chagas disease	00426 Wormian bones
00052 Bullous diseases of esophagus	00088 Cleidocranial dysostosis
00103 CREST syndrome	00103 CREST syndrome
00120 Diffuse esophageal spasm	00152 Esophagus: scleroderma
00002 Accessory spleen	00005 Acute pancreatitis
00664 Asplenia/polysplenia syndromes	00310 Pancreatitis
00092 Cold spleen lesion	00083 Chronic pancreatitis: surgical treatment
00192 Hot spleen	00309 Pancreatic pseudocysts
00377 Splenomegaly	00484 Zollinger–Ellison syndrome
00116 Dense spleen	00537 Echogenic pancreas
00235 Liver–spleen scan	00187 Hereditary pancreatitis
00003 Achalasia	00006 Adenoma sebaceum
00147 Esophageal dysmotility	00404 Skin lesions of tuberous sclerosis
00071 Chagas disease	00403 Tuberous sclerosis (Bourneville disease)
00337 Pseudo-achalasia	00367 Shagreen patches
00152 Esophagus: scleroderma	00450 Phakomatoses
00120 Diffuse esophageal spasm	00745 Acoustic neuroma
00858 Esophagitis	00488 Angiomyolipoma

DISCUSSION

The amount of information in the world often increases faster than our ability to organize it and make sense of it. In medicine, this problem of “information overload” is particularly acute. For example, more than 526,000 bibliographic citations were added in 2003 to MEDLINE, a collection indexed by the U.S. National Library of Medicine.⁶ Although information technology has improved our ability to publish and distribute information, it also has created challenges in finding, categorizing, linking, and effectively using the information made available. According to Malone et al.,⁷ there are three classes of filtering techniques: (1) cognitive, based on the content of documents; (2) social, or collaborative, which is based on human judgments; and (3) economic, based on the cost of producing or reading documents.

Content-based approaches are by far the most common way in which information resources are organized. Almost all Internet-based resources offer the ability to search by keywords, that is, a word or phrase that appears within a document or figure caption. Keyword searching has a number of problems. The item of interest may not use the exact word or phrase that is sought. Misspelled words—either in the source document or in the search string—may prevent relevant documents from being found. Although keyword searches are quite efficient computationally, they only identify strings of letters, not the underlying concepts. For example, a search for “gallstones” would not find a document that contained the term “cholelithiasis.” Conversely, a search for an extremely common term (e.g., searching for “CT” in an image database) might yield too many documents. Boolean and proximity operators can improve upon simple keyword searches for content-based filtering. Boolean searches combine search terms using Boolean-logic operators such as AND, OR, or NOT. Proximity searches look for words that occur close together within a document; typically, they must be separated by no more than three words.

To overcome the limitations of keyword searching, collections of documents are typically indexed by topics, which may be organized into a hierarchical ontology. In such indexing schemes—such as the American College of Radiology’s *Index for Radiological Diagnoses*⁸ and the National Library of Medicine’s Medical Subject

Headings (MeSH)—terms can be assigned hierarchical relationships. Concepts can be ordered hierarchically from the most general to most specific. Usually, an ontology will permit multiple relationships: thus, gallbladder carcinoma could be recognized as a specialization of the concepts “gallbladder disease” and “neoplasm.” The Unified Medical Language System (UMLS) Metathesaurus provides a large lexicon of medical terms, including the ACR *Index* and MeSH vocabularies, along with synonyms, related terms, and their hierarchical relationships.⁹ Developing an ontology or applying an existing one to index a collection’s content can be very labor-intensive.

Traditional collaborative filtering techniques are based on identifying user preferences and delivering recommendations based on those preferences. User-based CF algorithms require a database that records users’ preferences for the available items, and they utilize the entire user-item database to generate a prediction (Figure 3). First, they identify a subset of users, known as “neighbors,” who have a history of agreeing with the target user. Thus, neighbors are other users who either rate items similarly or tend to select (or buy or read) similar items. A variety of metrics can be used to measure similarity, including statistical measurements (e.g., the Pearson correlation coefficient) and probabilistic approaches (e.g., entropy-based uncertainty measures). Once a neighborhood of users is formed, a variety of techniques can be used to combine the preferences of neighbors to produce a prediction or “top-*N*” recommendation for the target user.¹⁰ Also known as nearest-neighbor or memory-based filtering, the user-based approach is the more popular and widely used form of collaborative filtering.

Neighborhoods can be selected using correlation thresholding, where all neighbors with absolute correlation values above a specified threshold are selected, or by choosing the “best *N*” correlates. It may happen that very few neighbors will have a high correlation, and thus it may not be possible to generate meaningful predictions for some items. Similarly, for the “top-*N*” approach, picking a large value of *N* may result in too much noise for those with high correlates, whereas picking a small *N* can cause poor predictions for users with low correlates.

Once a neighborhood has been selected, the predictions can be generated. Three techniques

Fig 3. Collaborative filtering (CF) process. (Adapted from Ref. 16.)

have been applied. First, one can compute the weighted average of user ratings using the correlations as the weights. This weighted average assumes that all users rate items with approximately the same distribution. Second, the weighted mean of all neighbors' ratings is computed. Rather than take the explicit numeric value of a rating, a rating's strength is interpreted as its distance from a neighbor's mean rating. This approach attempts to account for the lack of uniformity in ratings. Third, to account for the differences in the range of users' ratings, a "z-score" can be computed for each rating, which assigns significance to a rating as a function of the rating value, its difference from the neighbor's mean, and the degree of variance in the neighbor's ratings neighbors.¹⁰

Despite their successes, user-based recommender systems have two major limitations: sparsity and scalability.¹¹ The problem of sparsity arises in many settings where the amount of historical information for each user and for each item is often quite limited. For example, for an online bookstore with millions of users, it is very unlikely that any single user will have purchased even 0.01% of the 2 million available offerings. As a result, user-based CF systems cannot accurately compute the neighborhood and identify the items to recommend, which leads to poor recommendations. The second problem, scalability, arises because nearest-neighbor algorithms require computations that grow linearly with the number of users and items. When dealing with millions of users and millions of items, systems can suffer serious problems because of the number of computations to be performed. One way to reduce the computational complexity is to cluster

the users and to use the clusters to derive the recommendations. Although clustering can significantly speed up the computations, it tends to decrease the quality of the recommendations.

An alternative to a user-based CF system is one that bases its recommendations on the items selected.^{11,12} Because the bottleneck in conventional user-based systems is the search for neighbors among a large user population of potential neighbors, item-based algorithms can avoid this problem by first exploring the relationships among items to develop a model of users' preferences for each item based on preferences for similar items. In these approaches, the historical information is analyzed to identify relationships where the selection of one or more items often leads to the selection of another item (or a set of items). One can hypothesize that the relationships among items are relatively stable, which may provide the desired quality of recommendation while permitting the algorithm to scale with increasing collections of users and items. Item-based algorithms are up to 28 times faster than the traditional user-based recommender systems and provide recommendations of equal or greater quality.¹¹

Item-based algorithms first determine the similarities between the various items and then use that information to identify the set of items to be recommended. The key aspects of these algorithms are (1) the method used to compute the similarity between the items, and (2) the method used to combine these similarities in order to compute the similarity between a basket of items and an item to be recommended. Two different methods can be applied to compute the item-to-item similarity. The first method models the items

as vectors in the user space and uses the cosine function to measure the similarity. The second method computes the item-to-item similarity using a technique inspired by the conditional probability between two items, extended so that it can differentiate between users with varying amounts of historical information as well as differentiate between frequently and infrequently selected items.¹¹

Collaborative filtering offers a number of advantages over content-based filtering.¹⁰ Collaborative filtering takes advantage of the knowledge and opinions of people who previously have accessed documents within the collection. Collaborative filtering takes into account attributes such as quality, clarity, presentation style, and not just content. Whereas content-based approaches can be applied only to textual information (including text captions for images or other files), collaborative filtering can be applied to both textual and nontextual data such as images, sound, movies, and software.

Radiologists are developing large online collections of digital images, continuing medical education (CME) articles, video presentations, and other online resources for education and decision support. The Medical Image Resource Center (MIRC; <http://mirc.rsna.org>) was created under the aegis of the Radiological Society of North America to establish a community of Web-based libraries of imaging information. A central goal of MIRC is to enable professionals in radiology to create and publish teaching files, other educational materials, research data, and other materials more easily and to gain more convenient access to new and existing materials.¹³ Eurorad (<http://www.eurorad.org>) is a project led by the European Association of Radiology whose goal is to build a "huge and exhaustive case file of diagnostic and interventional radiology." Eurorad has published more than 1500 cases containing more than 7100 images.¹⁴ The Virtual Hospital (<http://www.vh.org>), a digital health sciences library created in 1992 at the University of Iowa, contains thousands of textbooks and booklets for health care providers and patients.¹⁵ Other large knowledge collections to which collaborative filtering can be applied include online journals and CME activities, such as RSNA's InteractED (<http://www.rsna.org>). User-based collaborative filtering can be applied to knowledge resources where users are

known (by login IDs or permanent Web-browser "cookies") and is well suited to subscription-based Web sites, such as online journals or collections of CME articles.

Collaborative filtering is a powerful tool by which users can benefit from the experiences of others. The approach can be applied to Web-based medical knowledge resources to improve a Web site's usefulness and ease of navigation. The recommendations obtained by item-based collaborative filtering increased the utilization of CHORUS documents. Collaborative filtering has found wide application in Internet-based commercial (e-commerce) sites. Whereas content-based approaches can be applied only to textual information (including text captions for images or other files), collaborative filtering can be applied to both textual and nontextual data such as images, sound, movies, and software. The current work is the first reported application of collaborative filtering in the medical literature.

CONCLUSIONS

Collaborative filtering represents one form of "mass customization," in which generalized information can be tailored to suit the needs of an individual. It improves a user's ability to locate information of interest: the paths followed by prior users become recommendations to those who follow. Collaborative filtering also reduces the burden faced by builders of large collections of images or text; the automatic nature of collaborative filtering can replace or augment the tedious task of manually applying an indexing scheme to the site's constituent items. As online radiology resources continue to grow in number and volume, it becomes increasingly important to organize them efficiently and retrieve information from them in a useful manner. Collaborative filtering is an established technique from the field of information retrieval that can be applied to improve navigation of online radiology knowledge resources.

REFERENCES

1. Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J: GroupLens: an open architecture for collaborative filtering of

Netnews. Proceedings of the ACM Conference on Computer Supported Cooperative Work, Chapel Hill, NC, pp 175–186, 1994

2. Hill W, Stead L, Rosenstein M, Furnas G: Recommending and evaluating choices in a virtual community of use. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, pp 194–201, 1995

3. Konstan J, Miller B, Maltz D, Herlocker J, Gordon L, Riedl J: GroupLens: applying collaborative filtering to Usenet news. *Commun ACM* 40:77–87, 1997

4. Shardanand U, Maes P: Social information filtering: algorithms for automating “word of mouth.” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Denver, CO, pp 210–217, 1995

5. Kahn CE Jr: CHORUS: a computer-based radiology handbook for international collaboration via the World Wide Web. *Radiographics* 15:963–970, 1995

6. National Library of Medicine Fact Sheet: The National Library of Medicine. U.S. Department of Health and Human Services, Public Health Service. <<http://www.nlm.nih.gov/pubs/factsheets/nlm.html>>. Accessed 6 July 2004

7. Malone TW, Grant KR, Turbak FA, Brobst SA, Cohen MD: Intelligent information-sharing systems. *Commun ACM* 30:390–402, 1987

8. American College of Radiology: Index for Radiological Diagnoses, 3rd edition, Reston, VA: American College of Radiology, 1986

9. Lindberg DAB, Humphreys BL, McCray AT: The unified medical language system. *Methods Inf Med* 32:281–291, 1993

10. Griffith J, O’Riordan C: Collaborative filtering. Technical Report NUIG-IT-160900. Department of Information Technology, National University of Ireland, Galway, Ireland, 2000. <<http://citeseer.ist.psu.edu/griffith00collaborative.html>>. Accessed 28 April 2004

11. Karypis G: Evaluation of item-based top-*N* recommendation algorithms. Department of Computer Science, University of Minnesota, Minneapolis, MN. Technical Report CS-TR-00-046, 2000

12. Deshpande M, Karypis G: Item-based top-*N* recommendation algorithms. *ACM Trans Inf Sys* 22:143–177, 2004

13. Siegel E, Channin D, Perry J, Carr C, Reiner B: Medical Image Resource Center 2002: an update on the RSNA’s Medical Image Resource Center. *J Digit Imaging* 15:2–4, 2002

14. Vorwek D: Neue internationale Wege in der radiologischen Fort- und Weiterbildung <http://www.eurorad.org>—ein Projekt der EAR zur Onlinepublikation radiologischer Daten. [New international networks in radiology graduate and continuing education: <http://www.eurorad.org>—an EAR project for online publication of radiological data]. *Radiologe* 42: 109–112, 2002

15. Galvin JR, D’Alessandro MP, Erkonen WE, Knutson DL, Lacey DL: The Virtual Hospital: a new paradigm for lifelong learning in radiology. *Radiographics* 14:875–879, 1994

16. Sarwar BM, Karypis G, Konstan JA, Reidl J: Item-based collaborative filtering recommendation algorithms. Proceedings of the Tenth International Conference on the World Wide Web, Hong Kong, pp 285–295, 2001