Conceptual Impact-Based Recommender System for CiteSeer$^{\mathrm{X}}$

Kevin Labille
Department of Computer Science and Computer Engineering
University of Arkansas
Fayetteville, AR 72701, USA
kclabill@uark.edu

Susan Gauch
Department of Computer Science and Computer Engineering
University of Arkansas
Fayetteville, AR 72701, USA
sgauch@uark.edu

Ann Smittu Joseph
Department of Computer Science and Computer Engineering
University of Arkansas
Fayetteville, AR 72701, USA
ann@email.uark.edu

ABSTRACT

CiteSeer^{x} is a digital library for scientific publications written by Computer Science researchers. Users are able to retrieve relevant documents from the database by searching by author name and/or keyword queries. Users may also receive recommendations of papers they might want to read provided by an existing conceptual recommender system. This system recommends documents based on an automatically-constructed user profile. Unlike traditional content-based recommender systems, the documents and the user profile are represented as concepts vectors rather than keyword vectors and papers are recommended based on conceptual matches rather than keyword matches between the profile and the documents. Although the current system provides recommendations that are on-topic, they are not necessarily high quality papers. In this work, we introduce the Conceptual Impact-Based Recommender (CIBR), a hybrid recommender system that extends the existing conceptual recommender system in CiteSeer^{x} by including an explicit quality factor as part of the recommendation criteria. To measure quality, our system considers the impact factor of each paper's authors as measured by the authors' h-index. Experiments to evaluate the effectiveness of our hybrid system show that the CIBR system recommends more relevant papers as compared to the conceptual recommender system.

Categories and Subject Descriptors

Information Systems [**Information retrieval**]: **Retrieval tasks and goals**:Recommender systems

General Terms

Performance, Reliability, Design, Experimentation

Keywords

Recommender System, h-index, Content-based Recommender System, CiteSeer^{x}, Information Retrieval

CBRecSys 2015, September 20, 2015, Vienna, Austria.

1. INTRODUCTION

In recent years, recommender systems have become ubiquitous, recommending movies, restaurants, and books etc. The recommendations ease information overload for users by pro-actively suggesting relevant items to the users, moving the burden of discovery from the user to the system. The number and type of applications that use recommender systems keeps growing [1]; one practical application that is of interest to researchers in any domain is the ability of recommender systems to suggest relevant scientific literature. These systems can expedite scientific innovation by helping researchers keep abreast of new publications in their fields and also help new researchers learn about the most important literature in an area new to them. Digital libraries can employ recommender systems that suggest papers to their users based on each user's research interests. However, an effective recommender system should not only consider the subject of a paper, it should also take into account the paper's quality when making recommendations. To this end, we present a recommender system that recommends scientific papers based on user preferences as well as paper quality as measured by the authors' impact factors to provide recommendations of high-quality papers that are relevant to the user's research area. To help CiteSeer^{x} users locate scientific papers related to their work, a citation-based recommender system was developed by Chandrasekaran et al. in 2008 [4] . Although citations are effective at identifying papers that have relevant content and are also high quality, this approach is only effective in recommending papers with many citations. These unfortunately tend to be older papers that have been published long enough ago to generate many citations. Especially in a fast-moving domain like computer science, researchers need to know about recent contributions to their field, yet recent papers have few citations. To solve this problem, a content-based recommender system for CiteSeer^{x} was developed by Pudhiyaveetil et al.[8]. This conceptual recommender system automatically builds conceptual profiles for users based on their interactions with the system. It also builds conceptual profiles for each document and recommends papers based on conceptual matches between document and user profiles. Even though the recommendations were shown to be more relevant than those produced by a keyword-based recommender system, they are not always high quality papers that the researcher wanted to read. Our objective is to improve upon the conceptual recommender system by providing better quality recommendations to the users. To do so, we developed a recommender system that recommends papers based on the paper authors' impact factors. We combined the impact-factor based recommendations with the concept-based recommendations in varying proportions to create a hybrid recommender system. We evaluated the effectiveness of the conceptual recommender system, the impact-factor recommender system, and the hybrid recommender system and found that the hybrid recommender system provides the most accurate recommendations. The rest of this paper is organized as follows: In section 2 we review related work. Section 3 describes the Conceptual Impact-Based Recommender (CBIR) system in detail. In section 4, we present our experimental evaluation to analyze the effectiveness of our recommender system. Finally, we present our conclusions and discuss future work in section 5.

2. RELATED WORK

The design of a recommender system can vary based on the nature of user feedback or the availability of data. There are three main approaches: collaborative filtering, content based recommender systems, and recommender systems that are a hybrid of the two [1]. The first approach generates recommendations based on similarities between the users' behavior or/and preferences. In contrast, content-based approaches recommend items to the users based on similarities between the attributes of the items themselves [10]. Collaborative approaches are typically used when semantic features cannot easily be extracted from the items, so indirect evidence based on user's likes or ratings must be compared. To be effective, collaborative filtering requires a large active user community to avoid the well-known ”cold-start” problem in which there are many more items to be recommended than there are users with likes or ratings upon which recommendations can be based. On the other hand, pure content-based recommender systems do not consider external information that might be available from the users, e.g., popularity. For these reasons, many recommender systems employ a hybrid approach combines both of the previously-described approaches.

Content-based recommender systems match the users' preferences to each items' features to recommend new objects [10]. Many share the approach of building a user profile from a set of features extracted from previously liked items. This user profile is then compared to the features of all items in the collection and the most similar items are recommended to the user [12]. This type of recommender system can be used in domain for which semantically relevant features can be extracted and it is particularly well-suited for domains that include textual items as scientific literature or domains with annotations such as movies or music [12]. Kompan et al. used this approach to recommend news articles on a web site [9]. In this domain, the volume of articles and the dynamic nature of news make collaborative filtering infeasible so they implemented a content-based recommender system based on cosine similarity that suggested articles that best matched an implicitly constructed user model [9].

Our work is a hybrid approach that enhances a content-based recommender system with a quality measure to recommend scientific literature. According to Beel et al., recommender systems for research papers are flourishing with more than 80 approaches existing today that have been discussed in over 170 articles and patents [2]. Such recommender systems are useful for researchers to be up to date in their research area. Many content-based recommender systems represent the user interests and the documents as weighted keyword vectors. One example is [13] in which $tf * idf$ weights are calculated for keywords and the cosine similarity measure is used to determine the relevancy of a paper to a user's profile. An approach similar to ours is used in [5]. In their work, each paper's features are represented as concepts created by automatically extracting keyphrases. User profiles are constructed from the concepts in previously viewed papers and the recommender system matches the user profile concepts to each papers' concepts to suggest new papers in a scientific library. In [8], a conceptual recommender system was presented that recommends research papers for CiteSeer^{x} users. Unlike the previous work, the concepts for each paper are assigned by automatically classifying papers into a set of concepts defined by a pre-existing ontology. A conceptual user profile is implicitly built as users view papers in the collection and this user profile is used to recommend conceptually similar papers.

The content-based recommender systems can recommend literature that is similar in topic to the user's profile, but it does not necessarily recommend high-quality papers. Although there is no perfect way to measure the quality of articles, the Impact Factor (IF) introduced in 1955 is still considered the best way to evaluate a paper's scientific merit [6]. There are several types of IFs, including the widely used h-index that evaluates a researcher's impact [7]. It has been recently used is several fields such as health services research [3], business and management [11] or even academic psychiatry [14] . Although the work in [5], [8], and [13] are similar to ours, our recommender system expands upon their work by incorporating a quality factor as measured by the authors' h-indexes.

3. APPROACH

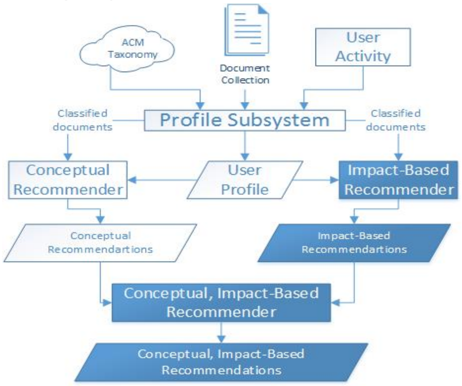

Figure 1: Architecture of the CIBR

The architecture of the Conceptual Impact-Based Recommender System (CIBR) is shown in Figure 1. The Profile Subsystem classifies all documents in the CiteSeer^{x} database into the 369 predefined categories in the ACM Computing Classification System (CCS). Documents manually tagged with ACM categories by their authors are used as the training set for a k-nearest neighbor classifier. As users interact with the system, the documents that they examine are input to the Profile Subsystem. The categories associated with each examined document are combined to create a weighted conceptual user profile. This user profile is used by both the Conceptual Recommender and the Impact-Based Recommender described in the following sections. The outputs of these two Recommenders are combined to produce the recommendations from the CBIR.

3.1 Concept-Based Recommender System

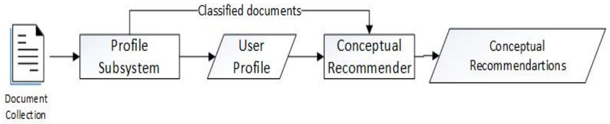

Figure 2: Conceptual Recommender System Architecture

As a user views documents in CiteSeer^x, the Profile Subsystem builds a conceptual user profile for them by accumulating the concept weights associated with the documents that the user examines. The Conceptual Recommender System then recommends documents to the user based on the similarity between each document's conceptual profile and the user's conceptual profile [8]. The weight of the conceptual match between document i and user j is calculated using the cosine similarity function over all M=369 concepts in the ACM taxonomy:

$$ConceptualWeight_{ij} = \sum_{K=1}^{M} (cwt_{ik} * cwt_{jk})$$

Where
cwt_{ik} = weight of concept k in document profile i and
cwt_{jk} = weight of concept k in user profile j as explained and detailed in [8].

3.2 Impact-Based Recommender System

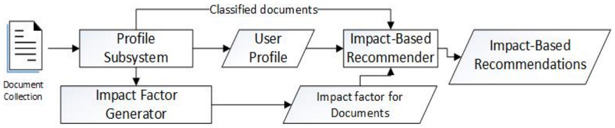

Figure 3: Impact-based Recommender System Architecture

The Impact Factor Generator precalculates an impact factor for each document in the collection as measured by its authors' h-indices. As described by Hirsch, an author has an h-index of m based on his/her N published articles if m articles have at least m citations each, and the other N-m articles have no more than m citations each [7]. The impact factor for a document is calculated by finding the h-index value of each of the authors of the document and then selecting the highest h-index value. Thus, document i's h-index is equal that of its most impactful author:

$$ImpactWeight_i = \max_{l \in A_{il}} (hindex_{il}) \tag{1}$$

Where
A_{il} = list of the authors l of document i
Since the impact factor is independent of users, the Impact-Based recommendations would be the same for all users, i.e., the most impactful documents in the entire collection. We do, however, use the user profile to filter out documents from categories in which the user has shown no previous input. Thus, Impact-Based Recommender returns high-impact documents from categories of some interest to the user. We tried other approaches to calculate the impact factor among which we consider the sum of each authors' h-indices. This particular method is limited since the highest weighted papers would usually be the ones with many authors.

3.3 Conceptual Impact-Based Recommender System

The Conceptual Impact-Based Recommender System (CIBR) combines the Conceptual Weights and the Impact Weights to produce its recommendations. The two sub-component weights are normalized to fall between 0 to 1 using linear scaling and then combined based on a tunable parameter, α. The weight of the conceptual impact match between document i and user j, γ_{ij}, is calculated using:

$$\gamma_{ij} = \alpha * C'_{ij} + (1 - \alpha) * I'_i \tag{2}$$

Where

$$C'_{ij} = \text{normalized } ConceptualWeight_{ij} = \frac{ConceptualWeight_{ij} - min_j(ConceptualWeight)}{max_j(ConceptualWeight) - min_j(ConceptualWeight)}$$

$$I'_i = \text{normalized } ImpactWeight_i = \frac{ImpactWeight_i - min_j(ImpactWeight)}{max_j(ImpactWeight) - min_j(ImpactWeight)}$$

α = controls the relative contributions of two sub-weights

By varying α from 0 to 1, we can adjust the relative contributions of two underlying recommender systems. When $\alpha = 0$, the CBIR is a pure impact-based recommender system whilst when $\alpha = 1$, the CBIR is a purely Conceptual recommender system.

4. EXPERIMENTAL EVALUATION

4.1 Subjects and Dataset

We conducted several experiments to measure the effectiveness of our hybrid recommender system. Experiments were done with 30 subjects, undergraduate and graduate computer science and computer engineering students from the university of Arkansas. We use the 2190179 documents in our snapshot of the CiteSeer^x, a digital library and a search engine for computer and information sciences literature. Because previous experiments have shown that profiles become stable after viewing 20 papers, users we asked to search for and view at least that many papers related to their own research area. Based on those documents, user profiles were automatically constructed for each user

4.2 Evaluation Method

The goal of this experiment was first to determine what combination the conceptual match and the paper quality is most effective in our hybrid recommender system. The relative combinations of the two is given by the equation in Section 3. By changing the value of α we are able to control the relative contributions of the two recommender systems with α = 0.0 being a pure impact-based recommender system and α = 1.0 being a pure conceptual recommender system and α = 0.5 using even contributions from both. We varied the value of α from 0.0 to 1.0 with an increment of 0.1 for each of the subjects in the experiment and for each value of α we collected the top ten recommended documents. For each

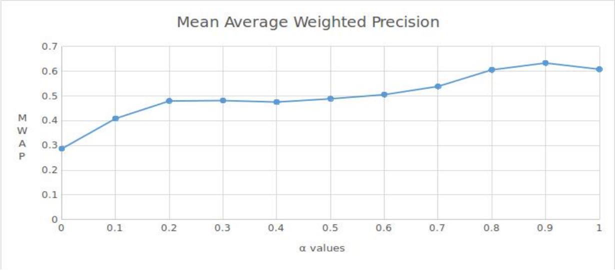

Figure 4: Mean Average Weighted Precision for every α

user, we presented them with the set of all documents recommended by any of the versions of the system (removing duplicates) in random order. They provided explicit relevance feedback by rating the papers as very relevant (2), relevant (1), or irrelevant (0). We then used the Mean Average Weighted Precision (MAWP) of each user for each α as a metric. The MAWP is essentially the Mean Average Precision modified to handle weights from 0..2 rather than just Boolean relevance judgments. The mean of every MAWP for each α is calculated and summarized in Figure 4. As shown on Figure 4, an α of 0.9 gives the best results, 0.6355, meaning that a 90% contribution from the conceptual recommender system and a 10% contribution from the impact-based recommender performed the best. For the second part of our analysis, we compared the effectiveness of the three recommender systems head-to-head. The hybrid recommender system with $\alpha = 0.9$ outperformed the conceptual recommender system's MWAP of 0.6083 ($\alpha = 1.0$) by 4.5% relative (or 2.72% absolute) and the impact-based recommender system's MWAP of 0.2867 ($\alpha = 0.0$) by 121.67% relative or 34.88% absolute. Both of these results are statistically significant ($p < 0.05$), based on the paired two-tailed student t-test.

5. CONCLUSION AND FUTURE WORK

In this paper, a hybrid recommender system was introduced that recommends high quality papers to CiteSeer^x users. The new recommender combines a conceptual recommender system along with an impact-factor-based recommender system. The former incorporates the user's preferences represented as a concept vector whilst the latter incorporates paper quality using the authors' impact factors as measured by their h-indexes. User experiments were conducted to compare the concept-based recommender system and the impact-based recommender system with our hybrid system. The results confirm that our hybrid recommender generates relevant documents as compared to the conceptual or the impact-factor-based recommender. Future work could consider using social networks of co-authors or differential weighting of the papers. Another direction would be to investigate the effectiveness of our hybrid recommender system by considering the g-index that gives a stronger weight to highly-cited papers as compared to the h-index. Alternatively, we could use the e-index that complements the h-index by distinguishing authors having the same h-index but different numbers of citations.

6. ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation grant number 0958123 : Collaborative Research: CI-ADDO-EN: Semantic CiteSeer^x

7. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[2] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, and A. Nürnberger. Research paper recommender system evaluation: A quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, pages 15–22. ACM, 2013.

[3] Y. Birks, C. Fairhurst, K. Bloor, M. Campbell, W. Baird, and D. Torgerson. Use of the h-index to measure the quality of the output of health services researchers. *Journal of health services research & policy*, 19(2):102–109, 2014.

[4] K. Chandrasekaran, S. Gauch, P. Lakkaraju, and H. P. Luong. Concept-based document recommendations for citeseer authors. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 83–92. Springer, 2008.

[5] D. De Nart and C. Tasso. A personalized concept-driven recommender system for scientific libraries. *Procedia Computer Science*, 38:84–91, 2014.

[6] E. Garfield. Journal impact factor: a brief review. *Canadian Medical Association Journal*, 161(8):979–980, 1999.

[7] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[8] A. Kodakateri Pudhiyaveetil, S. Gauch, H. Luong, and J. Eno. Conceptual recommender system for citeseerx. In *Proceedings of the third ACM conference on Recommender systems*, pages 241–244. ACM, 2009.

[9] M. Kompan and M. Bieliková. Content-based news recommendation. In *E-commerce and web technologies*, pages 61–72. Springer, 2010.

[10] P. Lops, M. De Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.

[11] J. Mingers, F. Macri, and D. Petrovici. Using the h-index to measure the quality of journals in the field of business and management. *Information Processing & Management*, 48(2):234–241, 2012.

[12] M. J. Pazzani and D. Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[13] S. Philip and A. O. John. Application of content-based approach in research paper recommendation system for a digital library. *International Journal of Advanced Computer Science & Applications*, 5(10), 2014.

[14] S. Selek and A. Saleh. Use of h index and g index for american academic psychiatry. *Scientometrics*, 99(2):541–548, 2014.