Sponsored Ad-Based Similarity:
An Approach to Mining Collective Advertiser Intelligence

Jessica Staddon
Palo Alto Research Center
staddon@parc.com

ABSTRACT

We present a method for mining the intelligence of advertisers to detect product similarities and generate accurate recommendations. In contrast to conventional recommendation algorithms, our approach is completely automated and relies solely on publicly available data, namely, the linking of advertisements with Web content. We present a general framework for leveraging linked advertisements to detect object similarity, and provide experimental evidence that the approach yields useful product recommendations.

1. INTRODUCTION

It's widely agreed that one of the core problems of information retrieval is detecting similarity between objects. Web page similarity supports useful ranking of Web query results (e.g. [10]), user preference similarity enhances recommendations (e.g. [8]) and similarity between text documents enables document categorization (e.g. [9]). Indeed, much research has focused on identifying new methods for measuring similarity (e.g. through hyperlinks or keyword overlap). An additional source for similarity information, that is largely untapped today, is online advertising. Online advertisers link the appearance of their advertisements to content they believe is likely to be generated by, or at least of interest to, potential customers. For example, sponsored links are attached to specific search engine queries and online articles on certain topics. Hence, the presence of the same advertisement with different pieces of content is an indication of similarity between the pieces of content.

In contrast to existing methods that infer similarity through object attributes that were created for other purposes, online advertisements are explicitly created *because* a similarity is perceived by the advertiser. For example, the furniture retailer, Pottery Barn, purchased the keywords "leather" and "ottoman" because it sees a high similarity between those terms and the interests of its customers[1]. In addition, the fact that these content-linked advertisements are the result of competition amongst advertisers suggests that, when present, advertisements are an accurate indication of similarity. Indeed, the number of businesses focused on helping advertisers pick their ad terms (see, for example, [2, 5]) attests to the effort that advertisers put into identifying terms highly similar to the interests of their customers.

While there is evidence that sponsored links enable accurate similarity detection, sponsored links alone are not enough for comprehensive similarity detection. The advertiser's goal is to attach keywords to ads to maximize returns on advertising investment, not to attract all potential customers. In practice this means that while objects A and B may be quite similar, they may not share any sponsored links. Hence, sponsored links are just one tool for detecting similarity, albeit a potentially powerful tool given the rapid and continued growth of keyword-based advertising.

We present a framework for mining collective advertiser intelligence to detect similar objects. Our framework builds on the simple observation that if object A and object B each lead to the display of sponsored ad C, then this is an indication of similarity between A and B. This observation leads to a simple method for generating commercial product recommendations. Recommender systems typically rely on extensive user input or analysis of the products considered for recommendation (for a survey, see [1]). In contrast, our approach is completely automated, requires no a priori knowledge of the brands and relies only on publicly available data (i.e. sponsored links). Hence, unlike most existing recommender systems, our approach does not disadvantage the new retailer who doesn't have a large database of customer transactions to mine, or the retailer who is introducing new products.

We provide experimental evidence that online sponsored ads can be used to extract relevant brand recommendations from a heterogeneous collection of brands. Our results indicate that even a large online retailer like `Amazon.com` may benefit from harvesting collective advertiser intelligence. For example, consider the recommendations `Amazon.com` makes under the heading, "Customers who viewed this item also viewed", for the stroller "BOB Revolution Duallie". It is difficult for the customer to find competing brands of stroller in the recommendations as most are for other BOB strollers (i.e. BOB strollers of different colors, or with slightly different features). In particular, of the top 11 `Amazon.com` recommendations, 7 are for BOB strollers. We applied the approach of this paper to a pool of 99 stroller brands (all of which are for sale through `Amazon.com`) and arrived at the ranked list of

[1]On May 6, 2008, Pottery Barn appeared as a sponsored link on Yahoo! for the query, "leather ottoman"

ADKDD '08 August 24, 2008, Las Vegas, Nevada, USA.

recommendations shown in Table 1. Despite knowing nothing about the strollers beyond their associated sponsored links the algorithm recommended 5 other 2-child strollers (as is the BOB Revolution Duallie) made by other manufacturers (i.e. not BOB) and 7 new brands in total. In contrast, the Amazon results include only 3 new 2-child strollers and 6 other BOB strollers. Given that the customer is already clearly aware of the BOB brand, the additional BOB recommendations seem of limited value.

OVERVIEW. This paper is organized as follows. In Section 2 we discuss related work. Our model and algorithm follow in Section 3. In Section 4 we present and analyze our experiments, and we conclude in Section 5.

2. RELATED WORK

We focus on recommendations as an application of our technique. Recommendations are a well studied area (for a survey and representative papers see [1]). Popular techniques for recommendations include mining a large database of user transactions and detecting similarity amongst users to leverage for predictions. In particular, most recommender systems take a collaborative filtering approach [4]. For example, customers are detected to be similar if they have rated items similarly, and purchases made by similar customers serve as recommendations (for surveys, see [6, 12]. Collaborative filtering can also be done at the good level, that is, goods are "similar" if they tend to be purchased together. This form of collaborative filtering is used by `Amazon.com` [8].

In contrast to collaborative filtering, our approach relies solely on publicly available information, rather than a private database of transactions, and requires no knowledge of the products under consideration. For these reasons it may be attractive to new retailers or for existing retailers who are introducing new products. That said, our technique for similarity detection can also naturally enhance an existing recommender system. For example, if a user has rated product A positively, and sponsored links indicate product B is similar, then B may be ranked more highly than other goods that are not known to be similar to A.

Viewed broadly, our work is another example of detecting similarity between Web objects (see, for example [3, 7]). Much work in this areas uses Web page hyperlinks to relate Web pages, and hence, our work can be seen as an extension of these ideas to advertising links. We are not aware of any work using linked advertisements to detect similarity.

3. MODEL

We consider objects that can be associated with each other through links they share to other objects. It can be helpful to represent these objects and associations as a graph. In particular, a representation that is compatible with [7] is a directed graph with nodes $\{O_i\}_i$ representing objects (e.g. sets of keywords) and advertisers (e.g. sponsored link URLs). A directed edge from O_i to O_j indicates that the advertiser represented by O_i has linked an ad to O_j (e.g. by buying the keywords associated with O_j).

Similarity between different brands is detected when advertisers associate ads with each brand. Advertisers may buy the keywords of the brands they sell (e.g. the advertiser `automall.com` buying the keywords "Kia", "Toyota", "Honda", etc.) or the brands of competitors (for example, on 5/7/2008, the search term "Kia" returned a sponsored link for Toyota (`www.ToyotaRetail.com`) on Yahoo!). In the case of the latter, there may be edges between advertisers as well.[2]

Section 4, focuses on recommendations of brands. We use the term "brand" to refer to a name that is associated with a collection of goods (e.g. "Toyota") and "product" to refer to a specific good (e.g. "Toyota Prius"). For example, Table 1 lists products, whereas, Tables 4, 3, 5 and 6, all contain brands.

We calculate recommendations by associating with each brand (product) a set of sponsored links. For a given brand (product) we rank the other brands (products) using a similarity measure [11], thus generating a ranked set of recommendations.

Our experiments use the Jaccard similarity measure (also known as the Jaccard index), hence we briefly repeat the definition here, for completeness. Let A and B be subsets of a set C. The Jaccard index of A and B is the ratio of the number of elements A and B have in common, to the number of elements in at least one of A or B:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Figure 1 is an example of a directed graph representing the relationships between sponsored links and brands with the associated Jaccard index values.

4. EXPERIMENTS

We manually created a collection of 350 brands of clothes, baby equipment (i.e. cribs, strollers, etc.), shoes and luggage. These brands were amassed from online "department stores" like Nordstrom (`http://www.nordstrom.com`) and Amazon (`http://www.amazon.com`). To define a "ground truth", we grouped the brands into the category that seemed to be most closely associated with the brand according to the results of search engine queries on the brands. Most of the brands naturally fell into a single category (e.g. "Graco" in baby equipment and "Ameribag" in luggage), however there were some cases of brands that branched into multiple categories, and for such brands we made a subjective decision based on what we saw as the brand's main product line.

The following 2 experiments were performed on this data set of 350 brands. They differ only in the transformations applied to the sponsored links associated with each brand. The purpose of the transformations was to explore the impact of more general advertisements (e.g. `www.amazon.com` vs `www.amazon.com/shoes`) on the recommendations. The specifics of the 2 experiments follow.

EXPERIMENT 1. For each brand we issued a single Google query on the brand name (e.g. "Graco") and scraped the resulting first page of sponsored links.[3] Each Google query was made through a URL formatted to be unlinked to any user account so that the choice of sponsored links was not overtly biased by user history[4] For example, the url for the "Graco" query is: `http://www.google.com/search?q=graco`.

From the scraped sponsored links we first excluded known aggregators and department stores (e.g. `Shopzilla.com`, `Shopping.com`, `Jcpenney.com`, `Amazon.com`, `Nordstrom.com`) since sponsored links from such sites could

[2]We don't exploit this possibility in our experiments.

[3]For almost all queries the first page contained all the sponsored links.

[4]This is not a fail-proof solution, as a user history could potentially be built up by cookies or IP address and this might influence the selection of sponsored links.

Source:	Amazon.com	Sponsored Link-based Section 3
1	BOB Revolution Duallie 12	BOB Revolution Duallie
2	InStep Safari Swivel Double Jogging	BOB Ironman
3	BOB Revolution Duallie	Baby Jogger City Mini
4	Schwinn Free Wheeler 2 Double Jogging	Baby Jogger City Elite
5	BOB Strides Fitness Duallie	Combi Twin
6	BOB Sport Utility	Bumbleride Rocket
7	Baby Jogger City Mini	Peg Perego Aria Twin
8	BOB Revolution Stroller	Bumbleride Flyer
9	BOB Revolution Duallie	Chicco Citta Twin
10	BOB Ironman Sport Utility Duallie	Zooper Tango
11	Graco Quattro Tour Duo	Dreamer Design Ditto Deluxe

Table 1: A comparison of the "Customers who viewed this item also viewed" recommendations on Amazon.com on May 13, 2008 with our recommendations generated from sponsored links. Our approach includes a wider variety of stroller brands and may be more useful, as the customer is already aware of BOB strollers.

potentially cause similarity to be detected between unrelated brands.[5] We summarize the brands in Table 2; the sponsored link data was gathered on 1/27/08.

We associated each brand name with the complete URLs of the scraped sponsored links, and then calculated the Jaccard index (Section 3) to measure similarity between the brands. The Jaccard index was chosen because the wide variety of brands created a very sparse similarity matrix and for such matrices more intuitive measures of similarity (e.g. Hamming distance) might generate results that are less easy to interpret. Recommendations are ranked by Jaccard index.

The data was quite sparse with an average of .05 shared sponsored links between different products. The average number of sponsored links per brand was 3.65 with a maximum of 10 sponsored links on the first page of hits for the brand Mountain Buggy. The number of shared links is only slightly higher within a category, indicating that there is not a large overlap amongst the keywords purchased by advertisers in a given category. Hence, the similarities detected when considering inputs from several advertisers are likely to be quite different than when focusing on a single advertiser.

A coarse metric for evaluating the usefulness of our recommendations comes from category matching. That is, if the recommendation falls into the same category as the initial brand (according to our hand-coding of the categories) then it is a good recommendation in some broad sense. Against this metric, our approach performed well. Out of the top recommendations for each brand with a Jaccard index of at least .125 (which yields 8 recommendations per product on average), 85% fell in the correct category (i.e. in the same category as the brand for which the recommendation was made), and more than 96% of the recommendations with Jaccard index at least .2 (which yields 2-3 recommendations per product on average) are correctly categorized. A graph comparing Jaccard values and accuracy is in Figure 2.

Since Figure 2 is essentially an evaluation of precision as a function of the Jaccard index, it is natural to also consider the complementary information retrieval measure, recall. In our setting, recall translates into the fraction of products in the correct category that are included in the recommendations, as a function of Jaccard index. Hence, if the recommendations include all brands with a Jaccard index of at least 0, then recall is 1 (all brands are recommended). For larger Jaccard values, recall drops off rapidly. For example, a minimum Jaccard index of $.125$ gives a total recall of 1.7%, and a minimum Jaccard index of $.0001$, gives a total recall of 8.3% (6.4% of clothes, 7.3% of baby equipment, 10% of luggage and 23.8% of shoes). This reflects the fact that many brands in each category share no sponsored links.

It is not immediately clear that a quickly declining recall is problematic for recommendations. A good recommender system filters out all the goods that aren't very similar to the available inputs (e.g. a good under consideration or the customer's preferences) and this translates into reduced recall. In addition, we emphasize that as the trend toward more online advertising continues, recall will consequently improve. In addition, recall and precision may be improved by using sponsored links associated with other search engines (e.g. Yahoo! or MSN) to detect similarity between brands.

We highlight some of the results of the experiment for four different types of brands: clothing, baby equipment, shoes and luggage.

CLOTHING. For the 171 brands that we hand-coded as clothing, the average number of sponsored links was 3.2 and the average number of shared links between distinct brands was .07. The most common "incorrect" recommendations were for luggage brands that included a large line of handbags (e.g. Elliott & Lucca). The maximum number of shared sponsored links between distinct clothes brands was 4 (e.g Marc Jacobs and Moschino) but there were only 4 such pairs of brands with such a large overlap. This category had some strongly dominant advertisers. For example, `www.designerapparels.com` purchased 33 distinct clothing brand names.

In Table 3 we show example clothing recommendations.

BABY EQUIPMENT. For the 99 brands that we hand-coded as baby equipment, the average number of sponsored links was 3.56 and the average number of shared links between distinct brands was .07. Miscategorized recommendations were very rare for baby equipment, probably due to the fact that brands

[5] We retained links from such sites provided they were specific enough, e.g. `http://amazon.com/shoes` was retained, but `http://www.amazon.com` was discarded.

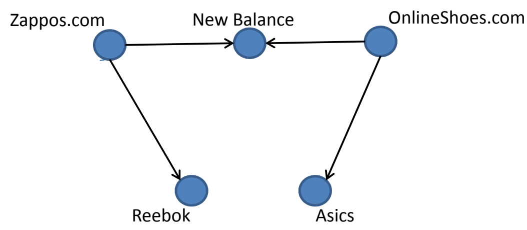

Figure 1: An representation of sponsored links and their associated brands in graph form. The Jaccard index for New Balance and Reebok is $1/2$ as is the Jaccard index for New Balance and Asics. However, J(Reebok, Asics)=0.

	Clothing	Baby Equipment	Shoes	Luggage	Total
Number of Brands	171	99	67	13	350
Average Number of Sponsored Links	3.2	3.6	5.1	2.9	3.65

Table 2: A summary of the inputs to our experiment on generating brand recommendations from sponsored links. Categorizations were created manually.

in this group were unlikely to have offerings in the other categories. In Table 4 we show example baby equipment recommendations.

Some popular brands were apparent. For example, the brand Zooper had 7 sponsored links and our algorithm used those links to detect similarity with 23 other brands. `BabiesRUs.com` was the largest advertiser in this category, with ads linked to 15 different brands in our experimental set of 350 brands.

SHOES. For the 67 brands that we hand-coded as shoes, the average number of sponsored links was 5.12 and the average number of shared links between distinct brands was .3. The most common miscategorized recommendations were for clothing and luggage brands as there are a lot of cross-over brands in each category (e.g. bag makers who also produce shoes). This category had the largest number of purchased ads by a single advertiser; `www.zappos.com` bought 39 brand names in our experimental set. In Table 5 we show example shoe recommendations.

LUGGAGE. For the 13 brands that we hand-coded as luggage, the average number of sponsored links was 2.9 and the average number of shared links between distinct brands was .07. Due to the small sample size there were not any additional useful statistics. In Table 6 we show example luggage recommendations.

EXPERIMENT 2. For the same data set of 350 brands and the same sponsored links (gathered on 1/27/08) we conducted a second experiment. First, we did not exclude known aggregators and department stores from the scraped sponsored links. Second, we truncated the sponsored links to their "root". More precisely, each sponsored link was reduced to the largest link ending in ".com". So, for example, the sponsored link `http://www.amazon.com/shoes` became `http://www.amazon.com`. This was done to generate more similarity data (i.e. more links between brands). This resulted in the truncation of approximately 32% of the sponsored links, however, it did not completely accomplish the goal of reducing the links to their core component. For example, there were a handful of advertisers who prepended their link with the brand, for example, `http://LizClaiborne.womenclothing.com` and `http://LLBean.womenclothing.com`. A more thorough approach would characterize all such variations and reduce them to the common portion of their url. That said, our naive approach succeeded in catching most of the brand-personalized sponsored links.

Somewhat surprisingly, the effect of these changes was quite modest. The precision of the results was quite similar to experiment 1, although the curve grew slightly less quickly (e.g. a precision of 94.5% for a Jaccard value of .3, as opposed to precision of 100% in experiment 1). In addition, recall improved somewhat, although perhaps not significantly. For example, in experiment 2, a Jaccard value of .1 gives a recall of 3% as opposed to 2.2% in experiment 1.

The continued high precision for the results suggests that the Jaccard similarity measure is serving to effectively filter out "generic" sponsored links like `amazon.com`, and hence,

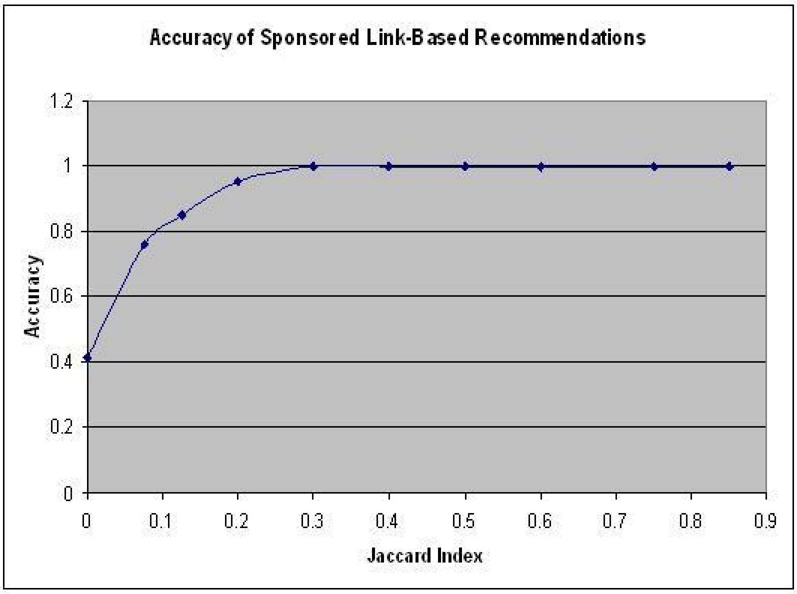

Figure 2: The percentage of correctly categorized recommendations (i.e. recommendations that are in the same category as the source they are based on) as a function of the minimum Jaccard index of the recommendation.

Recommendations for "Benetton"	Benetton	Club Monaco	Christian Dior	Pelle Pelle	Bisou-Bisou	Brioni	Elliott & Lucca
Jaccard Index	1	.5	.33	.286	.25	.25	.11
Category	Clothes	Clothes	Clothes	Clothes	Clothes	Clothes	**Luggage**
Recommendations for "Norma Kamali"	Norma Kamali	Bogner	Nicole Miller	BCBG	Ellesse		Elliott & Lucca
Jaccard Index	1	.167	.167	.1538	.133		.125
Category	Clothes	Clothes	Clothes	Clothes	Clothes		**Luggage**
Recommendations for "Sonia Rykiel"	Sonia Rykiel	Marc Jacobs	Paul Smith	Club Monaco	Moschino	Benetton	Prada
Jaccard Index	1	.273	.182	.167	.167	.143	.083
Category	Clothes	Clothes	Clothes	Clothes	Clothes	Clothes	**Luggage**

Table 3: Recommendations for three brands of clothing: Benetton, Norma Kamali and Sonia Rykiel. For each, the top 5 or 6 recommendations (ordered by Jaccard Index) and the first incorrect recommendation (i.e. mismatched category) is given.

Recommendations for "Maclaren"	Maclaren	Child Craft	Inglesina	Banana Fish	Halo EvenFlo	Innovations
Jaccard Index	1	.182	.154	.125	.111	.1
Category	Baby	Baby	Baby	Baby	Baby	Baby
Recommendations for "Valco"	Valco	Chicco	Clek	Inglesina	Baby Jogger	Silvercross Tike Tech
Jaccard Index	1	.2	.167	.167	.125	.125
Category	Baby	Baby	Baby	Baby	Baby	Baby
Recommendations for "Kettler"	Kettler	Chicco	Inglesina	Baby Jogger	Silvercross Tike Tech	Bumbleride
Jaccard Index	1	.182	.167	.143	.143	.133
Category	Baby	Baby	Baby	Baby	Baby	Baby

Table 4: Recommendations for three brands of baby equipment (i.e. strollers, cribs, etc.): Maclaren, Valco and Kettler. For each, the top 5 recommendations (ordered by Jaccard Index) are given. There were no incorrect recommendations (i.e. recommendations for another category of good) for any of the brands.

Recommendations for "Asics"	Asics	Seychelles	Ecco	Franco Sarto	Vans	Dr. Martens	Le Tigre
Jaccard Index	1	.2143	.2	.1875	.182	.176	.167
Category	Shoes	Shoes	Shoes	Shoes	Shoes	Shoes	**Clothes**
Recommendations for "Camper"	Camper	Enzo					
Jaccard Index	1	.125					
Category	Shoes	Shoes					
Recommendations for "Merrell"	Merrell	Easy Spirit	Franco Garmont	Skechers	Life Stride	Dance Now	
Jaccard Index	1	.143	.143	.133	.125	.125	
Category	Shoes	Shoes	Shoes	Shoes	Shoes	Shoes	

Table 5: Recommendations for three brands of shoe: Asics, Camper and Merrell. For each, the top recommendations (ordered by Jaccard Index) are shown. For Camper there are only 2 recommendations because of the small number of sponsored links. In the case of Asics the highest ranked incorrect recommendation (i.e. mismatched category) is shown; there were no incorrect recommendations for Camper and Merrell.

Recommendations for "Baggalini"	Baggalini	Damiro	A. Saks	Club Monaco
Jaccard Index	1	.182	.143	.11
Category	Luggage	Luggage	Luggage	**Clothes**
Recommendations for "Ameribag"	Ameribag			Olivier Strelli
Jaccard Index	1			.091
Category	Luggage			**Clothes**
Recommendations for "A. Saks"	A.Saks	Baggalini		La Baby
Jaccard Index	1	.143		.111
Category	Luggage	Luggage		**Baby**

Table 6: Recommendations for three brands of luggage: Baggalini, Ameribag and A.Saks. For each, the top recommendations (ordered by Jaccard Index) are shown, as well as the first incorrect recommendation (i.e. a recommendation for a non-luggage product). Due to the small pool of luggage products (13) there were a small number of recommendations. In particular, only 2 recommendations for A.Saks and just 1 for Ameribag (the trivial self-recommendation).

manually removing department stores and aggregators as we did in experiment 1, may be unnecessary. The small effect on recall may be an indication of the fact that the brand-personalized sponsored links were often used by specialty retailers (e.g. `http://BabyAge.com/Chicco`) and so reducing those links to their "root" did not generate a lot of new links between products.

5. CONCLUSION

We have introduced a new approach to detecting similarity between objects, based on linked advertisements. As the growth of online advertising continues, we believe this will prove to be an increasingly valuable approach that gets around the bootstrapping problem of new retailers without access to a customer database, and existing retailers who introduce new products. That said, our approach is certainly just one tool to detect similarity and it is best used in conjunction with other approaches.

There is substantial room for additional work on the broader question of how to use advertisements to detect similarity. One avenue that we suspect is particularly rich is using graph/set theory to derive additional information from the graphs resulting from the objects and their sponsored links.

In addition, the analysis of our approach is incomplete. A more extensive analysis involving more, and a wider variety of, brand types, and the use of additional search engines (e.g Yahoo!, MSN) to establish links, would be very valuable for evaluating the usefulness of the approach given the state on online advertising today.

Acknowledgement

The author is very grateful to Philippe Golle, Richard Chow and the anonymous referees for comments on an earlier version of this paper.

6. 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. IEEE Transactions on Knowledge and Data Engineering, 2005.

[2] adMarketplace. `http://www.ampkeywords.com`.

[3] D. Fogaras and B. Racz. Scaling link-based similarity search. WWW 2005.

[4] D. Goldberg, D. Nichols, B. Oki and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 1992.

[5] Google Advertsing Professionals. `https://adwords.google.com/select/professionalwelcome`

[6] J. Herlocker, J. Konstan, A. Borchers and J. Reidl. An algorithmic framework for performing collaborative filtering. *Proceedings of the SIGIR*, ACM, 1999.

[7] G. Jeh and J. Widom. SimRank: A Measure of Structural-Context Similarity. *KDD 2002*.

[8] G. Linden, B. Smith and J. York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, January-February 2003.

[9] M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[10] L. Page, S. Brin, R. Motwani and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Stanford University Technical Report, 1999.

[11] P. Tan, M. Steinbach and V. Kumar. *Introduction to Data Mining*, Addison-Wesley, 2005.

[12] H. Varian and P. Resnick. Recommender systems. *Communications of the ACM*, March 1997.