A Feature-Based Item Ranking Approach to Itemset Construction in Price Comparison Shopping Services

Kwanho Kim, Beom-Suk Chung, Yongsuk Yang, and Jonghun Park

Information Management Lab., Department of Industrial Engineering,
Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-744, Korea
{goalwisk, bumdo103, cncever4, jonghun}@snu.ac.kr

Abstract. Price comparison shopping services (PCSSs) are emerging as a convenient shopping tool. They provide a price-sorted itemset for a product of interest, which is a set of items constructed for some specific item type. Since itemset construction tasks in PCSSs often involve intensive labor, a method that can reduce the workload is of great importance. In this article, we propose an item ranking approach for itemset construction tasks, aiming to minimize itemset construction costs in PCSSs.

Keywords: Item Ranking, Itemset Construction, Price Comparison Shopping Services, e-Commerce

1 Introduction

In online shopping, price comparison shopping services (PCSSs) play an important role in enhancing shopping experience of customers by providing a price-sorted itemset for a product. An itemset is a set of items constructed for some item type [1]. Since the items in an itemset, called its member items, are from a single item type, customers are not only to able compare various purchase alternatives for an item type but also to reduce item search costs by eliminating the need of visiting individual shopping malls. A PCSS periodically gathers a huge number of new items from online shopping malls, and it constructs itemsets by assigning their memberships to itemsets.

Due to the massive volume of newly gathered items, itemset construction task, an essential part of PCSSs operations, requires a large amount of time and costs in general. This holds especially in case when human experts are required to be involved in the task. Therefore, an effective method that can suggest the membership likelihood of a newly gathered item with respect to an itemset is important for minimizing the workloads.

Although there has been some research effort related to the problem, most of the studies have relied mainly on detecting similar items based on duplicate detection methods rather than focusing on ranking items for itemset construction [2]. Moreover, the existing ranking methods based only on textual content may not produce satisfactory results due to non-consideration of item prices.

In this article, we propose an item ranking approach to itemset construction tasks that require human judgment to assign item memberships in PCSSs. In contrast to the

existing approaches, the proposed approach accounts for both the textual and price features of items, aiming to automating the itemset construction tasks by ranking newly gathered items with respect to a specific itemset according to their membership likelihoods for the itemset. It is expected that the proposed approach reduces itemset construction workloads by allowing human experts to focus only on top ranked items, obviating thorough investigation of the entire newly gathered items.

2 Term and Item Weighting Scheme

2.1 Term Weighting

Term weighting is an essential part for ranking items with respect to the itemset. In our approach, each term appearing in the descriptions of an itemset's member items is weighted based on its informativeness, cohesiveness, and appearance ratio. Then, top *K* weighted terms for an itemset are used for scoring items against the itemset with textual features. First, the informativeness of a term, which measures the importance at itemset level, represents how well it uniquely describes an itemset. We adopt the normalized inverse frequency to measure term informativeness [3], and the informativeness of term *t* is calculated as

$$inf(t) = \left(\frac{\log \frac{sf(t)}{(1)} + 1}{\log(N+1)_j} \right)^{\backslash 2}$$

where $sf(t)$ is the number of itemsets that consist of one or more member items containing term *t* in their descriptions, and *N* represents the number of itemsets.

Next, term cohesiveness, which is designed to capture term co-occurrence patterns in item descriptions, is presented. It was empirically observed that the descriptions of the member items for an itemset tend to contain the same or similar patterns of term co-occurrences, since the descriptions of those items from the same item type often include some specific phrases describing product codes and manufacturers. Specifically, the degree of cohesiveness of term *t* for itemset *S* is defined as

$$coh(S,t) = \frac{\left| \bigcup_{i \in /(S,t)} bag(i) \right|}{\sum_{i \in /(S,t)} bag(i)} \quad (2)$$

where */(S, t)* is the subset of the itemset S's member items that contains term *t* in their descriptions, *bag(i)* is a function that returns the bag-of-terms for the description of item *i,* and *lbag(.)* represents the number of distinct terms in a given term bag.

Lastly, we consider a term's importance based on its appearance ratio in the member items of an itemset compared to its non-member items. The appearance ratio of a term for an itemset becomes higher, if it intensively appears on the descriptions of its member items. We define *ar(S, t)* to be the appearance ratio of term *t* for itemset

S. In particular, two existing methods, bi-normal separation (BNS) [4] and information gain (1G) [5], are employed for computing the appearance ratio.

By mixing up the above three measures for weighting a term for a specific itemset, the final term weighting function is defined. The weight of term *t* for itemset S, $A_{s,t}$, is defined as

$$= inf\,(t) \bullet coh(S,\ t) \bullet ar(S,t). \qquad (3)$$

2.2 Price based Item Weighting

Along with the textual scoring of items, we attempt to give weights to items based on their prices against each itemset. Since the prices of the member items of an itemset are highly likely to be similar to each other, an item is less likely to be a member of an itemset if its price is far from those of the member items of the same itemset. Specifically, an item's weight is measured by using a non-parametric outlier detection method that detects the instances significantly different from the other instances [6]. The item weight of item *i* against itemset *S* is computed by

$$P s = \frac{/(s)v\ \mathrm{E}p_{iP}}{\mathrm{E}\,\mathrm{E}_{id(s)_{\backslash\cdot};\ el\ (S)} — P i''} \qquad \textbf{(4)}$$

where *I(S)* represents the member items of itemset *S*, and *p*, is the price of item *i*. Finally, based on the proposed term and item weighting methods, the score of item *i* against itemset *S*, which represents the membership likelihood of the item with respect to the itemset, is calculated as

$$score(S, i) = \{ \underset{s,t}{El}\, p_{5,t} \qquad (5)$$

where Q_s is the set of selected terms for scoring items against itemset *S*.

3 Experiment Results

We conducted experiments to show the effectiveness of the proposed approach by using a real-world dataset collected from Best Buyer (http://www.bb.co.kr) which is one of the top three PCSSs in Korea. We randomly selected 1.5K itemsets and 100K items, and the selected items were from 142 different online shopping malls The half of those items was considered as newly gathered items that require their memberships to be defined. The member items for each itemset had been manually constructed by human experts, and candidate member items for an itemset had been previously

filtered based on the product code corresponding the itemset before manual investigation of their memberships.

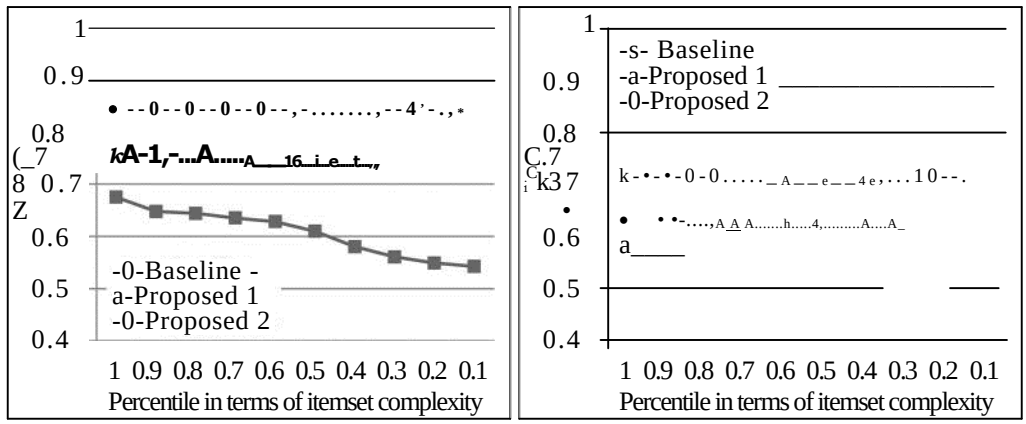
For each itemset, we selected top *K* weighted terms by utilizing Equation (3), and each item's weight for the itemset was obtained by Equation (4). Subsequently, the newly gathered items were ranked against an itemset by using Equation (5). The item ranking performances were measured by a modified version of normalized discount cumulative gain (NDCG) [7]. The modification is to address the effects caused by the filtered items through measuring the difference between the best and worst NDCG values.

Table 1. NDCG results of the item ranking performances according to the number of selected terms for scoring items, *K*. The performance improvements compared to the respective baselines are presented in parentheses.

K	BNS			IG		
	Baseline	Proposed 1	Proposed 2	Baseline	Proposed 1	Proposed 2
5	0.641	0.774 (+21%)	0.859 (+34%)	0.585	0.606 (+4%)	0.720 (+23%)
10	0.658	0.782 (+19%)	0.861 (+31%)	0.583	0.611 (+5%)	0.731 (+25%)
15	0.671	0.783 (+17%)	0.863 (+29%)	0.585	0.611 (+4%)	0.731 (+25%)
20	0.675	0.795 (+18%)	0.864 (+28%)	0.586	0.651 (+11%)	0.742 (+27%)

Table 1 shows the performance comparison results for the considered metric, and the results imply that the proposed item ranking method without item weighting, represented as proposed 1, can select better terms to rank items than the existing appearance ratio based method, denoted as baselines. On average, the performance improvements achieved by utilizing the proposed term weighting method were 18.52% and 5.98% compared to those obtained by utilizing the baselines: BNS and IG, respectively.

Moreover, the further performance improvements were observed when the proposed item weighting method (denoted as proposed 2) that considers item prices was additionally applied for ranking items. This means that the proposed item weighting method is effective to filter out outlier items in terms of their prices for an itemset. The improvements achieved by using the proposed method, which utilized both the term and item weighting methods, were 30.37% and 25.01% compared to those obtained by using the baselines: BNS and IG, respectively.



(a) BNS (b) IG

Fig. 1. Comparisons of the performances according to itemset complexities under K = 20. Figures (a) and (b) represent the performances when the appearance ratio measure is BNS and IG, respectively.

To show the robustness of the proposed approach, we also investigated the performances by using the subsets of the selected itemsets according to itemset complexities, as shown in Fig. 1. Here, the itemset complexity of an itemset represents the degree of difficulty in distinguishing its member items from its nonmember items, which is based on the entropy of the terms appearing on the descriptions of its member items [8]. In the figure, vertical and horizontal axis represents NDCG results and percentile in terms of itemset complexities, respectively.

Fig. 1 illustrates that the item ranking performances observed by using one of the baselines alone rapidly decrease as the itemset complexity is increased, while the performances obtained by utilizing the proposed methods show quite impressive results for all the cases. The performance differences between the results obtained by using the proposed methods and the baselines tended to become greater as the itemset complexity is increased. These observations suggest that the proposed methods are more robust in ranking items particularly for complex itemsets than the existing weighting methods we considered.

4 Conclusions

We presented a novel item ranking approach for itemset construction by considering term weighting which consists of term informativeness, cohesiveness, and appearance ratio, and price based item weighting. The experiment results show that the proposed approach provides satisfactory item ranking results. This implies that it can much reduce the manual workloads to itemset construction compared to the existing methods, resulting in cost reduction for constructing itemsets in PCSSs. In the future, we plan to enhance the method by exploiting sequential patterns of term occurrences and price changes.

Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea Government (MEST) (Nos. 2011-0004423 and 2011-0030814), and partly by Engineering Research Institute at Seoul National University and Interpark INT.

References

1. Bock, G.W., Lee, S.Y., Li, H.: Price Comparison and Price Dispersion: Products and Retailers at Different Internet Maturity Stages. International Journal of Electronic Commerce 11, 101--124 (2007)
2. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering 19, 169--178 (2007)
3. Valle, E.D., Ceri, S., van Harmelen, F., Fensel, D.: It's a Streaming World! Reasoning upon Rapidly Chaining Information. IEEE Intelligent Systems 24, 83--89 (2009)
4. Forman, G.: An Expensive Empirical Study of Feature Selection for Text Classification. The Journal of Machine Learning Research 3, 1289--1305 (2003)
5. Yang, Y., Pedersen, J.: A Comparative Study on Feature Selection in Text Categorization. In: 14th International Conference on Machine Learning, pp. 412--420. Morgan Kaufmann Publishers Inc., San Francisco (1997)
6. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying Density-Based Local Outliers. In: ACM SIGMOD International Conference on Management of Data, 29, 93--104 (2000)
7. Jarvelin, K., Kekalainen, K.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems 20, 422--446 (2002)
8. Ho, T.K., Basu, M.: Complexity Measures of Supervised Classification Problems. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 289--300 (2002)