PromotionRank: Ranking and Recommending Grocery Product Promotions Using Personal Shopping Lists

PETTERI NURMI, ANTTI SALOVAARA, and ANDREAS FORSBLOM, Helsinki Institute for Information Technology
FABIAN BOHNERT, Monash University
PATRIK FLORÉEN, Helsinki Institute for Information Technology

We present PromotionRank, a technique for generating a personalized ranking of grocery product promotions based on the contents of the customer's personal shopping list. PromotionRank consists of four phases. First, information retrieval techniques are used to map shopping list items onto potentially relevant product categories. Second, since customers typically buy more items than what appear on their shopping lists, the set of potentially relevant categories is expanded using collaborative filtering. Third, we calculate a rank score for each category using a statistical interest criterion. Finally, the available promotions are ranked using the newly computed rank scores. To validate the different phases, we consider 12 months of anonymized shopping basket data from a large national supermarket. To demonstrate the effectiveness of PromotionRank, we also present results from two user studies. The first user study was conducted in a controlled setting using shopping lists of different lengths, whereas the second study was conducted within a large national supermarket using real customers and their personal shopping lists. The results of the two studies demonstrate that PromotionRank is able to identify promotions that are considered both relevant and interesting. As part of the second study, we used PromotionRank to identify relevant promotions to advertise and measure the influence of the advertisements on purchases. The results of this evaluation indicate that PromotionRank is also capable of targeting advertisements, improving sales compared to a baseline that selects random advertisements.

Categories and Subject Descriptors: H.3.3 [**Information Systems**]: Information Search and Retrieval; H.4.2 [**Information Systems**]: Information Systems Applications

General Terms: Algorithms, Measurement, Applications

Additional Key Words and Phrases: Ranking, recommender Systems, personalization, retailing, advertising, user study

ACM Reference Format:
Petteri Nurmi, Antti Salovaara, Andreas Forsblom, Fabian Bohnert, and Patrik Floréen. 2014. PromotionRank: Ranking and recommending grocery product promotions using personal shopping lists. ACM Trans. Interact. Intell. Syst. 4, 1, Article 1 (March 2014), 23 pages.
DOI: http://dx.doi.org/10.1145/2584249

This work was supported by the Finnish Funding Agency for Technology and Innovation TEKES under the project Adaptive Interfaces for Consumer Applications (AICA). The work was also supported in part by the ICT program of the European Community, under the PASCAL2 network of excellence, ICT-216886-PASCAL2. The publication only reflects the authors' views.

Authors' addresses: P. Nurmi and P. Floréen, Helsinki Institute for Information Technology HIIT, PO Box 68, FI-00014 University of Helsinki, Finland and Helsinki Institute for Information Technology HIIT, PO Box 15600 FI-00076 Aalto University, Finland; A. Salovaara, Helsinki Institute for Information Technology HIIT, PO Box 15600 FI-00076 Aalto University, Finland; F. Bohnert, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia.

The reviewing of this article was managed by associate editor Pearl Pu.

DOI: http://dx.doi.org/10.1145/2584249

1. INTRODUCTION

Shopping is one of the most fundamental everyday activities. The way people shop is increasingly diversifying as conventional means of shopping are supplemented with other forms of shopping, such as Internet shopping and mobile commerce. Regardless of the mode of shopping, promotions play a central role in shopping [Lal and Matutes 1994; Lohse and Spiller 1998; Vakratsas and Ambler 1999]. Among other things, customers use promotions for budgeting and planning, whereas retailers use promotions to accelerate purchase cycles, stimulate sales of complementary products, motivate the use of loyalty cards, and attract new customers [Blattberg et al. 1995; Kumar and Leone 1988; Mauri 2003; Thomas and Garland 1996; Walters 1991].

The relevance of promotions that are presented to the user has a significant influence on their effectiveness [Hupfer and Grey 2005; Rettie et al. 2005; Wang et al. 2002]. Furthermore, when the presented promotions are considered irrelevant, they are easily perceived as spam and cause negative attitudes toward advertising in general [Barwise and Strong 2002; Dahlén et al. 2003; Tsang et al. 2004]. Techniques and means for improving the relevance of promotions are thus essential for achieving high levels of customer satisfaction. This is particularly the case when the promotions are presented on mobile devices, where limited screen size and cumbersome input capabilities reduce the possibilities of reaching the customers at the correct time.

The present article focuses on identifying relevant promotions in a mobile retailing context. We present PromotionRank, a technique for generating a personalized ranking of promotions based on the contents of the customer's personal shopping list. PromotionRank combines information retrieval techniques with recommender systems to tackle the following design challenges: (1) natural language shopping lists often contain spelling errors, colloquialisms, ambiguities, and so forth, making it difficult to establish a mapping with product categories; (2) customers often buy substitute brands compared to what they have noted on their shopping list [Kumar and Leone 1988]; (3) customers usually buy a substantial amount of other items than those they have noted on their shopping list [Thomas and Garland 2004]; and (4) customers appreciate information about promotions of relevant, less frequently purchased products, as observed by our experiments.

We evaluate the effectiveness of PromotionRank through offline experiments and two user studies. The offline experiments have been conducted using 12 months of anonymized shopping basket data from a large national supermarket and evaluate the individual phases of PromotionRank. The two user studies, on the other hand, focus on evaluating the performance of the overall PromotionRank technique. The first user study was conducted in a controlled setting using seminatural shopping lists and considers the relevance and interestingness of the promotions identified by PromotionRank for shopping lists of different lengths. The second user study was conducted within a large-scale supermarket with actual customers and their personal shopping lists. This study evaluates the capability of PromotionRank to identify relevant personalized advertisements in a real-world context. The results of the evaluations demonstrate that PromotionRank can identify relevant and interesting promotions based on limited user input provided by the personal shopping lists of the users, and that the performance of PromotionRank is not sensitive to the length of the shopping lists that are given as input. The results of the second user study demonstrate that these findings carry over to a real-world context and that PromotionRank is capable of targeting personalized advertisements, improving sales compared to a baseline that selects random advertisements.

Summary of Contributions

The main contributions of the article are summarized as follows:

—We demonstrate that items on the customer's personal shopping list, despite their limited and restricted content, can be used as input for personalization and to identify promotions that are relevant and interesting.
—We describe PromotionRank, a technique for identifying relevant and interesting promotions based on the user's shopping list.
—We demonstrate that recommendation techniques can be used to expand the original user input in a way that enables identifying a larger set of relevant content.
—We extensively evaluate PromotionRank, both in simulated and real-world settings, demonstrating that PromotionRank (1) can identify promotions that are both relevant and interesting, (2) is not sensitive to the length of the shopping list that is given as input, and (3) is capable of targeting relevant personalized advertisements in real shopping contexts.

2. RELATED RESEARCH

Different types of recommendations have been explored in the retailing domain [Kim et al. 2002; Linden et al. 2003]. A popular task has been to generate personalized product recommendations based on customer transaction data. For example, Cumby et al. [2004] predict customers' shopping lists using individualized classifiers that are trained using their recent transaction histories. The predictions are used to remind customers of forgotten products and to target promotions. The best predictions result from combining a top-N predictor and a simple discriminative classifier such as Winnow, Perceptron, or C4.5. Adomavicius and Tuzhilin [2001] discover association rules for individual customers and propose a validation layer for experts to validate the discovered rules. Demiriz [2004] uses association rules to find interesting item pairs and then rank the items using memory-based collaborative filtering. Mild and Reutterer [2001] compare the accuracy of memory-based collaborative filtering algorithms using different proximity measures, sparsities, and amounts of data. Algorithms seem to perform robustly with different amounts of data, but their accuracy decreases as sparsity increases. Lawrence et al. [2001] use customer and product class profiles to generate product recommendations about previously not purchased products. Field tests showed revenue increases of approximately 2%.

Instead of relying on customer transaction data, we have in our previous work [Nurmi et al. 2009a] explored the use of individual items on the customer's personal shopping list to recommend additional products. The recommendations were generated using a combination of information retrieval techniques and generalized association rules. A user study demonstrated that the approach generates product recommendations that customers consider interesting. Note that this technique is not applicable to the task considered in the present article, as it focuses on individual items and, contrary to the present work, assumes the user input is a complete expression of the customer's interests. Vindevogel et al. [2005] use scanner data to demonstrate that some of the captured associations are between mutually substitutable products. Therefore, the authors propose to use multivariate time series data to derive cross-price elasticities between products in order to determine effective promotional strategies, but similar techniques can also be used to generate product recommendations. Brijs et al. [1999] employ cross-price analysis, association rules, and a micro-economic model to identify positive cross-sale effects. Their technique helps to determine which items to have on sale.

Our work differs from previous work in three main aspects: (1) instead of using personal transaction history, our technique uses personal language shopping lists as the input to recommendation techniques; (2) instead of stimulating additional purchases through product recommendations, we attempt to stimulate additional purchases by recommending products that are promoted by the retail chain, enabling us to make recommendations that are beneficial to the customer within the actual supermarket environment (definitions for relevance and interestingness are given in Section 5.3); and (3) instead of focusing solely on determining relevant and accurate recommendations, we aim to identify promotions that are also considered interesting. The motivation for using personal shopping lists instead of the customer's transaction history is twofold. First, customer transaction data is highly sensitive, and data protection legislation in many countries forbids the linking of purchase information with the identity of an individual customer. Accordingly, accessing this information is challenging and only possible for retailers. Second, while the transaction history of a customer reflects his or her overall interests, contrary to his or her personal shopping list, it does not provide clues about the customer's current interests and needs. The motivation for focusing on existing promotions instead of product recommendations is that results from a user study that investigated customers' preferences regarding features in an intelligent mobile grocery assistant indicate that customers are more interested in information about relevant and actual special offers than suggestions for additional products [Bhattacharya et al. 2012]. Finally, the motivation for considering both relevance and interestingness is related to the way promotion strategies are designed; see Section 4.4.

3. BACKGROUND

Before introducing the PromotionRank technique, we describe the dataset that was used to construct the algorithm, as well as a grocery retrieval engine that we have constructed in our previous work [Nurmi et al. 2008a, 2008b].

3.1. Dataset

PromotionRank has been constructed using an anonymized shopping basket dataset that has been provided to us by a large Finnish supermarket chain. The data contains all purchases between January 2007 and February 2008. Overall, the data contains 12.4 million individual purchases of 18, 000 different products from one large supermarket. The purchases belong to 1.3 million different shopping baskets (i.e., receipts) and were made by 160,000 different customers. For each product, the data also contains a category name. The categories are based on the product hierarchy used by the supermarket chain, and there are a total of 923 different categories.

3.2. Grocery Retrieval

The way people normally write shopping lists contrasts with the way grocery stores maintain information about their products. Stores tend to use structured formats, for example, product-category hierarchies (or taxonomies), that contain formal language, whereas customers tend to use natural language for describing items. A typical handwritten grocery list can contain everything from generic item descriptions (e.g., `milk`, `juice`) to very specific items (e.g., a specific package of washing powder) [Keaggy 2011]. To overcome this discrepancy between the customers and the retailers, PromotionRank utilizes a grocery retrieval engine that takes items on the user's personal shopping list (e.g., `coffee`) as input and returns a ranked list of products that match the query.

The retrieval engine was constructed using the anonymized shopping basket dataset described previously. From each transaction in the dataset, we extracted the product and the corresponding category in the product hierarchy used by the supermarket chain. From this information we created a database of product and category names. From the database, we created two inverted indexes, one for words occurring in the product names and one for words occurring in the category names. Since our database has been constructed solely from purchase history information, we have very little textual information available, making the retrieval task nontrivial to solve.

Consider the task of mapping a query for an arbitrary shopping list item s into potentially relevant products. When the retrieval engine receives the query, it first translates the query into lowercase and extracts all package size- and unit-related parts (e.g., 100g or 1.5l) from the input. Next, a set of language-specific preprocessing steps are performed on the input, before the query is matched against the index. If the query fails to provide any results, a misspelling correction algorithm is applied on the query. As our work focuses on a Finnish retailing context, the engine currently operates in Finnish. However, translating the retrieval engine to other languages merely requires modifying the module responsible for carrying out the language-specific preprocessing steps.

In our case, the language-specific preprocessing consists of two steps. First, to overcome variations caused by different cases, we use lemmatization on the input query. We carry out the lemmatization using a freely available Finnish lemmatizer that is part of the Voikko-project[1] and based on the Suomi-Malaga grammar.[2] Second, we apply compound splitting on the input in order to split compound words into their constituents. The use of compound splitting is motivated by experiments at the monolingual track of the Cross-Language Evaluation Forum (CLEF), which have shown compound splitting to significantly improve the performance of Finnish retrieval [Airio 2006; Hollink et al. 2004]. We carry out the compound splitting using a variant of the algorithm of Monz and de Rijke [2002]; see Nurmi et al. [2008b] for details. Both the lemmatization and compound splitting steps are also applied during the construction of the retrieval index. As an example of how to adapt our system to other languages, translating our system to use English language merely requires applying stemming instead of lemmatization on the input words. In case of English, no compound splitting is required.

Once the query has been processed, we match it against the index used by the retrieval engine. We rank the products that match with the query using a probabilistic ranking framework that combines product purchase frequency with textual features. The ranking function $g(z, s)$ that we use approximates the conditional probability $p(z|s)$ of candidate product z being relevant given the query s. Formally, we consider the following approximation:

$$\begin{aligned} p(z|s) &= p(s|z)p(z)/p(s) \\ \Leftrightarrow \log p(z|s) &= \log p(z) + \log p(s|z) - \log p(s) \\ &\approx \log p(z) + \lambda BM25(z,s) + K \\ &\overset{\text{def}}{=} g(z,s) + K. \end{aligned} \tag{1}$$

Here $K = -\log p(s)$ is a query-specific constant that does not depend on the candidate product and that can be ignored in the ranking. The prior probabilities of candidate products $p(z)$, on the other hand, correspond to purchase frequencies, which can be modeled using a multinomial distribution with a Dirichlet prior. Finally, the function $BM25(z, s)$ corresponds to the value of the BM25 Okapi function for a query term s given a product z, which is defined as follows [Robertson et al. 1995]:

$$BM25(z,s) = \sum_{j \in s} \log \frac{N - n_j + 0.5}{n_j + 0.5} \frac{(\nu + 1) f_j}{f_j + \gamma((1-b) + bL)}. \tag{2}$$

Here, n_j is the number of product names that contain word j, f_j is the term frequency of word j (i.e., the number of times the word j appears in the product name z), and

[1] http://voikko.sourceforget.net.

[2] http://joyds1.joensuu.fi/suomi-malaga/suomi.html.

N is the total number of products. The variable L is the normalized document length, that is, the length of the current item divided by the average item length (in words). Finally, ν, b, and γ are predefined constants; see Robertson et al. [1995] for details and Nurmi et al. [2008a, 2008b] for the parameters that are used with the retrieval engine.

In our derivation, we have relied on the fact that the BM25 Okapi ranking function can be interpreted as a log-odds probability [Robertson and Walker 1994; Sparck Jones et al. 2000a]. The motivation for using the BM25 ranking method is that TREC evaluations have shown the model to perform well in various retrieval tasks [Sparck Jones et al. 2000a, 2000b]. Moreover, we prefer to use a robust and extensively tested model instead of developing new retrieval principles. In our case, the documents are relatively short, which means that there is basically no repetitive structure in the documents. As a consequence, a pure text search does not work effectively as it is not able to separate between the rare and the more popular products. This is why we need to consider also the prior probabilities of the items.

Shopping lists mainly contain generic product descriptions instead of referring to specific products. This can cause problems since the generic names also are likely to match nonrelevant products. For example, the query `milk` also returns products matching `milk chocolate`. However, products containing the term `milk chocolate` (typically) belong to the category `chocolates`, whereas products containing only the term `milk` belong to the category `milk products`. Since the category name `milk products` matches the original query, we can improve search results by boosting the scores of the matching products in this category. We utilize the product category information in the dataset to boost the rank of products that match both in product and category. Specifically, we use a weighted extension of BM25 that replaces the term and document frequencies in Equation (2) with weighted linear sums [Robertson et al. 2004]:

$$n'_j = dm_j + n_j \quad \text{and} \quad f'_j = dc_j + f_j, \tag{3}$$

where d is a weight term, m_j is the number of category names that contain word j, and c_j is the term frequency of word j calculated from the name of category c.

During the process of constructing our retrieval engine, we conducted a small-scale query analysis using 99 shopping lists that were collected from our target supermarket. This analysis revealed that around 5% of the shopping list items contain misspellings. Most of the misspelled words had only a single mistaken character, and none had more than two mistakes. Motivated by this result, the retrieval engine uses a distance index to overcome misspellings. The distance index contains word pairs and the distance between the two words. As the distance measure we use edit distance, which is defined as the minimum number of string operations (insertions, deletions, or substitutions) that are needed to transform one string into another [Crochemore and Rytter 2002]. Only word pairs with edit distance at most two are stored in the distance index. Thus, the distance index provides support in situations where the words contain one or two misspelled characters. The distance index is consulted whenever the original query does not return any results. Specifically, we fetch word candidates with edit distance one from the distance index and then repeat the query using these candidates, ignoring query words that are less than three characters long. If this query does not return any results, we repeat the same process for words within edit distance two. If this query also fails to return results, the retrieval system returns an empty result set.

4. PROMOTIONRANK

Our proposed technique, PromotionRank, consists of four phases: (1) candidate pool creation, (2) candidate pool expansion (CPE), (3) rank score calculation, and (4) ranking of promotions. For our experiments, we also consider a variant of PromotionRank that does not utilize the candidate pool expansion phase. The relationships between the

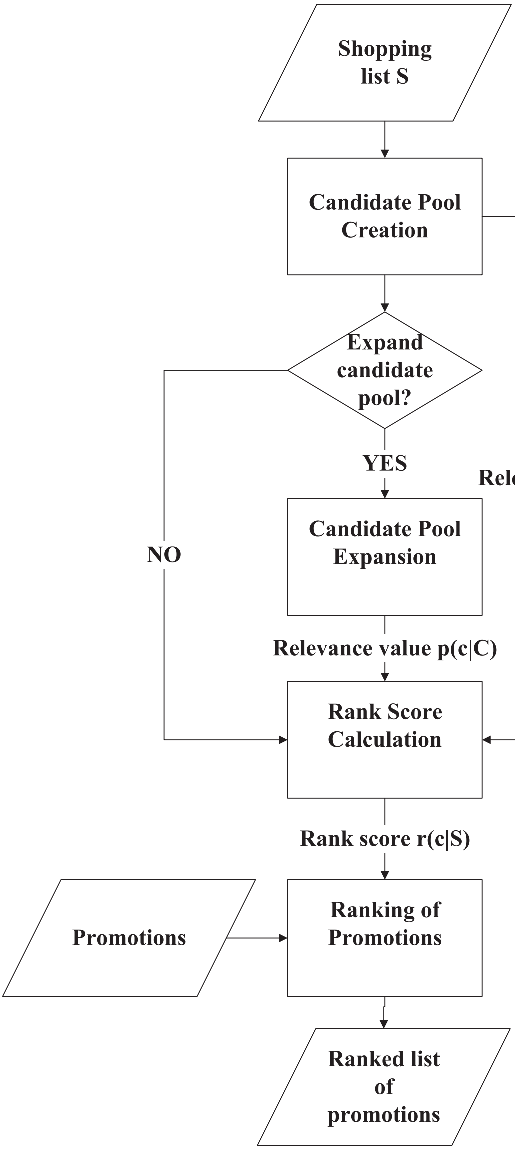


Fig. 1. Overview of the different phases in PromotionRank.

different phases are illustrated in Figure 1. In the following, we discuss each of the phases in more detail.

4.1. Shopping List Entry

The input to PromotionRank consists of the personal shopping list S of a customer. Without loss of generality, we assume the shopping list is given as text. In our experiments, the shopping lists are inputted manually using mobile text entry. Other possible ways to add items to the shopping list include scanning barcodes, extracting shopping lists from SMS messages, or extracting items with OCR from camera images.

4.2. Candidate Pool Creation

In the first phase of PromotionRank, the items on the customer's personal shopping list are mapped into relevant product categories. The mapping is performed on a category level because customers often consider substitute brands within the same category when sufficiently good promotions are available [Kumar and Leone 1988]. The product categories are then used to create a pool of candidate categories C (hereafter *candidate pool*) that serves as input to the subsequent phases.

To create the candidate pool, we use the grocery retrieval engine described in Section 3.2 to retrieve the top five matches for each shopping list item. The motivation for considering only five retrieval results is based on our user evaluations of the retrieval engine, which have indicated that further results are rarely perceived as relevant and can in fact decrease the quality of the candidate pool. As an example of how the candidate pool creation works, the shopping list `juice, milk` would issue two queries, one for `juice` and one for `milk`. The query `milk` returns different products and adds the corresponding categories such as `skimmed milk` and `99% fat-free milk` to the initial candidate pool, whereas the query `juice` would result in categories such as `orange juices` and `pineapple juices` being added.

Once the categories are added to the candidate pool, we calculate a relevance score for each of the categories that are added. The relevance scores are used in PromotionRank's last phase to determine the final ranking of the available promotions. The rank scores (i.e., the values of Equation (1)), are not directly comparable as they approximate probabilities up to a query-specific constant, which is intractable to estimate. Consequently, we need to normalize the rank scores before we can compare them across items. For a given query s, let z_q denote the product at rank q. We normalize the rank score of the product z_q using

$$\tilde{g}(z_q, s) = \frac{\exp g(z_q, s)}{\sum_{q=1}^{k} \exp g(z_q, s)}, \tag{4}$$

where k is the number of products that the retrieval engine returns (i.e., in our case, $k = 5$) and $g(z_q, s)$ is given by Equation (1). This normalization effectively corresponds to approximating the cumulative distribution function for the given item. As the values that our retrieval engine returns correspond to logarithms of probabilities that have a negative sign, an exponential function is used to ensure that the most relevant results have the highest score. Let Z_c denote the set of products that belong to category c. The probability of category c being relevant given the shopping list S is approximated using the maximum of the normalized scores, that is,

$$p(c|S) \approx \max_{s \in S, z \in Z_c} \tilde{g}(z, s). \tag{5}$$

Accordingly, the more relevant a product is to at least *one* of the items on the customer's shopping list, the more likely we assume it to be relevant.

4.3. Candidate Pool Expansion

Shopping lists rarely contain all items that a customer actually purchases [Thomas and Garland 2004]. To capture relevant items that are not included on the customer's shopping list, the second phase of PromotionRank, candidate pool expansion (CPE), supplements the candidate pool with additional product categories that are potentially relevant to the customer, resulting in an *expanded candidate pool* C'.

The intuition behind our approach to expand the candidate pool is to add product categories that are often purchased together with the categories that are already in the candidate pool. The expansion algorithm augments the candidate pool C with product categories $x \notin C$ for which the probability $p(x|C)$ is high. As shopping basket data is typically very sparse,[3] it is difficult to reliably estimate $p(x|C)$ from observed shopping data. To address this issue, we consider two alternative ways to approximate these probabilities: association rule–based CPE and probabilistic CPE.

Association Rule–Based CPE. Our first method expands the candidate pool with categories x that appear often with one of the categories c in the candidate pool C; that is, we consider the following approximation:

$$p(x|C) \approx \max_{c \in C} p(x|c), \tag{6}$$

where $p(x|c)$ is the confidence of the association rule $\{c\} \Rightarrow \{x\}$. Accordingly, an efficient way to apply this approximation is to consider association rules that are sorted by confidence. In our case, we use category-level association rules that contain a single item in their head and body. We examine the rules in descending order of confidence and add categories from matching rules to the candidate pool. We continue the expansion until either the number of categories in the candidate pool has tripled or the confidence of the rules falls below a predefined threshold; see Section 5.2 for the parameter values that were used in our experiments.

Probabilistic CPE. Our second method considers the probability of a category appearing together with all of the categories in the candidate pool. To make the calculation tractable, we make the simplifying assumption that the probability of a category x appearing together with category $c \in C$ is independent of the other categories in the candidate pool. We also assume that the probabilities of the different categories appearing in the candidate pool are identically distributed. These assumptions enable us to approximate $p(x|C)$ as follows:

$$p(x|C) \approx \prod_{c \in C} p(x|c) \Leftrightarrow \log p(x|C) \approx \sum_{c \in C} \log p(x|c). \tag{7}$$

We estimate the probabilities $p(x|c)$ using a multinomial distribution with a Dirichlet prior:

$$p(x|c) = \frac{p(x,c)}{p(c)} = \frac{n_{x,c} + \alpha}{(1+\alpha)\, n_c}, \tag{8}$$

where $n_{x,c}$ denotes the number of times that the categories x and c appear together in the data, and n_c denotes the number of times that the category c appears in the data. The variable $\alpha > 0$ is a smoothing constant that ensures the probabilities are nonzero when two categories do not appear together in the data; see Section 5.2. To expand the candidate pool, we rank the categories according to the probabilities given by Equation (7). Similarly to our first method, we then add new categories to the candidate pool until either the overall number of categories has tripled or the probabilities fall below a predefined threshold; see Section 5.2.

4.4. Rank Score Calculation

The third phase of PromotionRank assigns a rank score for all categories in the expanded candidate pool C'. The simplest way to accomplish this would be to utilize the overall purchase frequency of the corresponding product categories. However, this strategy is often suboptimal both for the retailer and the customer. Retailers often use

[3] In our dataset the sparsity of the data on the category level (i.e., the percentage of zero entries in the user-category matrix), is 98.98%.

frequently purchased products as loss leaders that are sold at small or even at negative margins [Lal and Matutes 1994]. Promoting frequently purchased products could thus result in small profit margins for the retailer. Loss-leader products are typically advertised prominently in local media and outside the store entrance, which implies that customers have likely been exposed to these promotions before entering the store. Customers also tend to have sufficient price knowledge for recognizing promotions of frequently purchased products even without explicit advertising [Vanhuele and Drèze 2002]. Instead of assigning a high rank to frequently purchased products, PromotionRank identifies and assigns a high rank to promotions of products that the customer considers both relevant and interesting.

We calculate the rank score of a category using the strength of the statistical dependency between the category and the candidate pool. We measure the strength of statistical dependency using the interest (or lift) between the category and the candidate pool (i.e., how much more likely the category is to appear together with the other categories in the candidate pool than independently of them). However, the sparsity of shopping basket data makes it impossible to reliably compare the category against the entire candidate pool. To address this, we use the maximum interest between the category x and any category $c \in C'$ to measure the interest between x and the candidate pool C', that is,

$$I(x, C') = \max_{c \in C'} \frac{P(x, c)}{P(x)P(c)}. \tag{9}$$

An efficient way to apply this criterion is to capture category-level association rules and to assign the interest of a category by going through the resulting rules in decreasing order of their interest.

The interest values $I(x, C')$ assume that all categories in the expanded category pool C' match the items on the customer's shopping list S equally well. This assumption rarely holds in practice. Instead of relying solely on the interest values, we integrate the interest values with the relevance values of the candidate pool creation and CPE phases to derive an overall rank score for the categories in the expanded candidate pool. We calculate the rank score of a category c given the shopping list S using:

$$r(x|S) = \begin{cases} p(x|S)\, I(x, C') & x \in C \\ p(x|C)\, I(x, C') & x \in C' \backslash C. \end{cases} \tag{10}$$

Both the relevance value $p(x|S)$ resulting from the candidate pool creation phase (Equation (5)) and the relevance value $p(x|C)$ resulting from the CPE phase (Equations (6) or (7))[4] can be interpreted as approximate probabilities that indicate how likely it is that the given category is relevant to a customer. Higher values of $p(c|S)$ indicate a higher likelihood of the category c being relevant to the customer's shopping list, whereas higher values of $p(c|C)$ indicate a higher likelihood of the category that is added to the candidate pool being relevant. Accordingly, $p(c|C)$ and $p(c|S)$ can be interpreted as discount terms that integrate uncertainty with the interest values given by Equation (9). Typically, the relevance values resulting from the candidate pool creation phase are significantly larger than the relevance values resulting from the candidate pool expansion phase. Consequently, PromotionRank tends to emphasize categories related to the items on the shopping list. As the results of the experiments demonstrate, this strategy results in a good balance between relevance and interest.

[4] Note that $p(x|C) = 1.$ whenever $x \in C$.

4.5. Ranking of Promotions

The ranking of promotions can be accomplished by first mapping each promotion item to its category and then ranking the promotions according to the rank scores of the corresponding categories. Promotions that do not match any of the categories in C' (or C when CPE is not utilized) remain last in the ranking. Ties are resolved using the purchase frequencies of the product categories. The resulting ranking can also be used to recommend personalized promotions (e.g., by selecting the top k promotions from the resulting personalized ranking).

5. EXPERIMENTS

In this section, we evaluate PromotionRank. We quantitatively evaluate the second stage of the recommendation process, candidate pool expansion, and report on the results from two user studies that evaluate the usefulness of the overall approach. The first phase of PromotionRank, candidate pool creation, has been evaluated using two user studies that have been reported in our previous work [Nurmi et al. 2008a, 2008b]. For completeness of presentation, we also briefly summarize these studies.

5.1. Candidate Pool Creation

We have evaluated the performance of the grocery retrieval engine that is responsible for the candidate pool creation phase in two user studies. Both user studies were carried out using a web interface that showed shopping list items together with the 10 topmost retrieval results for each item. Participants were asked to indicate the relevance of each result using a thumbs-up or -down button.

Our first user evaluation was carried out at university premises and resulted in 7,454 relevance assessments. The second study was conducted within a supermarket environment and resulted in 3,029 subjective relevance assessments. As part of the second study, we also compared our system against Lemur, a text-based search engine that uses BM25 without product frequencies for ranking the results. The results of the two evaluations indicate 80% to 85% accuracies for the two topmost ranks, and 70% to 75% accuracies for ranks three to five; we refer to Nurmi et al. [2008a] for more details about the first evaluation and Nurmi et al. [2008b] for more details about the second evaluation.

5.2. Candidate Pool Expansion

As the first step of the evaluation, we use the shopping basket data described previously to evaluate the two alternative CPE methods. We also compare the methods against a baseline that expands the candidate pool with the most frequently purchased product categories that do not appear in the candidate pool. In the experiments, we use each method to expand the candidate pool with 50 additional product categories.

As part of early experiments, we also explored an item-based collaborative filtering algorithm that uses conditional probabilities to estimate similarity between items [Deshpande and Karypis 2004]. This algorithm is widely used in many online recommender systems, and thus provides a good proxy for the performance of collaborative filtering techniques in general. Interestingly, this approach performed worse than our other methods, including the baseline, and was highly dependent on the choice of scaling parameters. For these reasons, we omit the results obtained with this algorithm. We posit that more generally, standard collaborative filtering algorithms—and other more advanced techniques, such as those based on latent variable models [Bellogin et al. 2011; Cremonesi et al. 2010; Koren 2008]—may be unsuited for the CPE phase due to their usage of symmetric similarity measures. This is because the usage of symmetric similarity measures effectively corresponds to scaling the probabilities of candidate categories with the inverse of their occurrence probabilities, resulting in frequently occurring categories having less influence than rarely occurring ones. While this is desirable in most recommendation scenarios [Deshpande and Karypis 2004; Cremonesi et al. 2012], it is not suited for the CPE phase, where the explicit goal is to identify categories that are likely to be missing from the customer's shopping list, regardless of their overall occurrence probability. Instead, the CPE phase benefits from using an asymmetric similarity measure. Note that association rules may be viewed as item-based collaborative filtering that utilizes an asymmetric similarity measure. This makes association rules well suited for our task.

Procedure. We evaluated the performance of the candidate pool expansion methods by randomly splitting the available shopping basket data into separate training and test sets. To decrease the influence of seasonal patterns on the results, the test set was generated by randomly selecting 7 days from each calendar month and including all purchases from these days in the test set. The purchases from the remaining days were used to train the expansion methods. This procedure was repeated 10 times and the results were averaged over the 10 repetitions. To simulate real shopping lists, which typically contain fewer items than what the customer actually purchases [Thomas and Garland 2004], the evaluation was conducted by randomly removing a fraction of products from each shopping basket in the test set. The different methods were then used to predict the product categories of the removed products. For each of the 10 test sets, we carried out the evaluation three times, removing either 30%, 50%, or 70% of products from each of the shopping baskets in the corresponding test set. To ensure that sufficiently many products remained after removal, only baskets with at least 10 purchases were considered. Each test set contained approximately 127,000 baskets and the average number of products in the baskets was approximately 20. We then used all three methods to create a ranked list containing 50 product category predictions for each basket in the test set. These predictions were compared to the categories of the withheld products and we used the results to calculate precision, recall, and F_1-score of the method at different ranks. The values were calculated separately for each test set and removal percentage pair, and we averaged the values over all of the shopping baskets in the corresponding test set.

Association Rule–Based CPE. The association rule–based CPE method was evaluated by mining association rules from the training dataset using a minimum support threshold of 0.01% and a minimum confidence threshold of 5%. These parameter values resulted in 49,423 rules between the different categories. The threshold on the minimum support was selected empirically considering the coverage of the rules and the total number of rules. Higher minimum support values rapidly decrease the coverage of the rules. The minimum confidence threshold, on the other hand, was set to avoid capturing weak dependencies between product categories. Note that since the association rule–based CPE ranks the candidate categories by confidence, the minimum confidence threshold has a negligible effect on the overall results of the method, as long as the number of rules that have been mined is sufficiently high.

Probabilistic CPE. The performance of the probabilistic CPE method, given by Equation (7), depends on the value of the smoothing parameter α. To determine the optimal value of α, we conducted the analysis with different values of α and selected the value that resulted in the best tradeoff between precision and recall. Generally smaller values of α resulted in better prediction performance in terms of F_1-score. Decreasing the value of α below 0.05 did not have any statistically significant effects on the resulting F_1-scores, and for this reason $\alpha = 0.05$ was selected for the final evaluation.

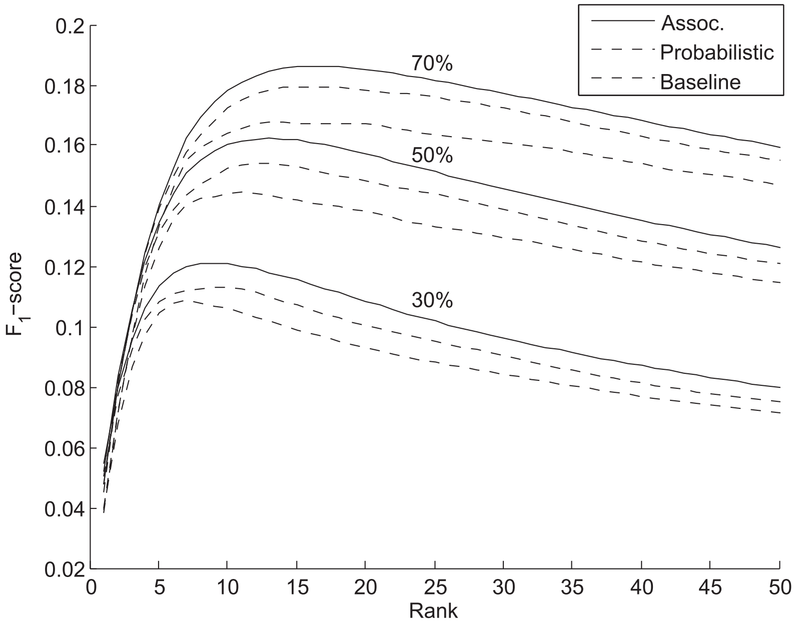


Fig. 2. F_1-score for the different methods with 30%, 50%, and 70% of the items removed.

Results. Figure 2 shows the resulting F_1-score for the different methods as a function of rank with 30%, 50%, and 70% of the items removed. Both of the CPE methods are better than the baseline, with the association rules performing slightly better. For example, the mean F_1-score over the 50 first ranks reached by the methods with 50% of items removed was 0.126, 0.136, and 0.142 for the baseline, probabilistic, and association rule methods, respectively. The results of the two CPE methods were approximately equal for the top five ranks. Despite the slightly higher F_1-score for the association rule–based expansion methods, the two user evaluations described in the next section were conducted using the probabilistic CPE method as it had slightly better precision.

5.3. User Evaluation

To evaluate the usefulness of the entire PromotionRank algorithm, we have conducted a user study that assessed how well the algorithm identifies relevant promotions. We evaluated the performance of PromotionRank with and without the CPE phase (PR-CPE and PR, respectively). As a baseline, we considered a system that maps each item of the customer's shopping list onto a product category and returns promotions whose categories match any of these categories, ranking them in order of purchase frequency.

Stimulus Data Generation. To make the evaluation setup as realistic as possible, we generated the stimulus material for the study using the 99 authentic shopping lists collected earlier at a large national supermarket. A post box had been installed at the supermarket's exit from December 2007 to January 2008 with a sign asking the customers to leave anonymously their shopping lists in the box. The following processing steps were performed on the lists to prepare the data for analysis. First, as the lists were collected around Christmas time, 12 lists that contained one or more products related to the Christmas season (e.g., candles) were removed. Next, nongrocery products (e.g., "shampoo"), idiosyncrasies ("food for our baby"), and ambiguous items ("some food”) were removed from the remaining 77 lists. In the third step, items were removed that were among the retail chain’s 50 most frequently purchased grocery products. This was done to reduce noise and increase the lists’ informational value. Fourth, different versions of each list were generated by first randomly drawing (without replacement) 2, 6 and 10 items from every sufficiently long list and then adding two random items from the list of the 50 most frequently purchased products. After this step was done, 69, 35, and 13 shopping lists were available for evaluations with 4, 8 or 12 items in each, respectively. Finally, of these lists, $13 + 13 + 13$ shopping lists were drawn (without replacement) so that the three different shopping list lengths (4, 8 and 12) were represented with 13 stimulus lists each. Three of the authors inspected the stimulus lists that were generated. Lists that contained repeated item categories (e.g., “milk” and “skimmed milk” in the same list) were regenerated and reinspected.

Promotions. In order to present the study participants with realistic promotions, we crawled available special offers from the web page of our partner supermarket chain over a 3-month period. The information that was collected included the name of the promoted product, the picture and price of the product, and the validity period of the promotion. In the user study, we included all promotions regardless of their validity period. To ensure that the attractiveness of promotions did not influence the participants’ assessments about the relevance of individual promotions, we removed the price and picture of the promoted products. Consequently, only the names of the promoted products were shown to the participants. We used our grocery retrieval engine to map the promoted products into product categories.

To summarize, we generated a corpus of realistic shopping lists (lengths 4, 8, and 12 items) and used them as input to generate promotion recommendations, which were evaluated with a sample of participants. These participants were different from those who had written the original lists. As the amount of promotions constrains the possible recommendations and as there typically is only one promotion per category, using third-party evaluators decreases the influence of personal preferences, such as brand preference or idiosyncratic shopping preferences, on the participants’ relevance assessments.

Procedure. The evaluation was carried out using a web page that visualized one stimulus shopping list at a time and showed a ranked list of five promotions for the given stimulus list. The recommendations were generated using one of the three methods: baseline, PR, and PR-CPE. Each participant was shown one list at a time, and he or she was asked to evaluate the shown promotions along two dimensions:

Relevance. How well the promotion fits the list presented. For instance, does the product belong to the ingredients that often go together in a certain meal?

Interestingness. How inspiring the promotion is given the shopping list. Would it work as an impulse purchase or provide an interesting alternative to the typical product of choice?

Responses were elicited using a 5-point Likert scale that was anchored at “do not agree” (=1) and “agree” (=5). Participants were instructed that answering “1” meant a poor suggestion that did not fit the shopping list, “3” meant a suggestion that was thinkable but not characteristic of the products in the list, and “5” meant a very good suggestion. “Cannot answer” was also available in case the suggested product was unfamiliar to the participant.

In each evaluation, both the ranking method and the stimulus list were selected randomly. The same list was never used more than once with each participant. The

Table I. Mean and Standard Deviation of Ratings for the Different Methods
Column PR refers to PromotionRank without candidate pool expansion, and column PR-CPE refers to PromotionRank with candidate pool expansion. The columns labeled *N* correspond to the number of items for which promotions were returned.

Rank	Relevance			Interestingness			N		
	Baseline	PR	PR-CPE	Baseline	PR	PR-CPE	Baseline	PR	PR-CPE
1.	3.48 (1.60)	3.97 (1.40)	3.82 (1.49)	2.96 (1.43)	3.44 (1.42)	3.39 (1.32)	54	59	57
2.	3.10 (1.62)	3.41 (1.55)	3.44 (1.48)	2.73 (1.39)	2.95 (1.47)	3.25 (1.18)	51	56	57
3.	3.09 (1.60)	3.35 (1.34)	3.14 (1.56)	2.98 (1.36)	3.08 (1.35)	3.28 (1.21)	46	48	57
4.	3.47 (1.67)	3.45 (1.33)	2.82 (1.55)	2.94 (1.35)	3.03 (1.10)	3.23 (0.98)	34	33	57
5.	3.44 (1.48)	3.21 (1.41)	2.77 (1.64)	3.41 (1.48)	2.96 (1.52)	3.28 (1.31)	32	24	57

participants were not informed about the method that was used to create the ranking of the promotions.

Participants were recruited from two university cafeterias at campuses of natural and behavioral sciences. The sample ($N = 38$) consisted of 23 females and 15 males (61% and 39%, respectively). For each completed evaluation page, a participant was rewarded with a confection. After completing a minimum of three pages, the participants were invited to continue if they wished to do so: 76% of the participants contributed the minimum and one participant provided 14 evaluations. In total, we collected 170 evaluations (4.5 evaluations per participant on average). There were no differences between male and female respondents in the number of lists evaluated.

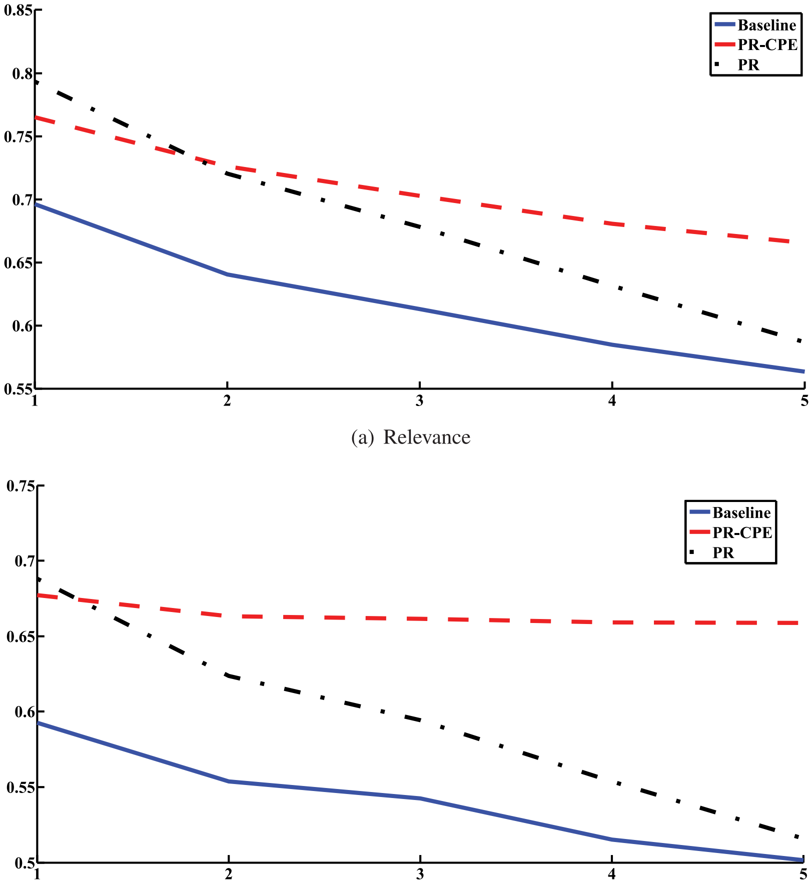
Evaluation Measures. As our main evaluation measure, we considered the normalized discounted cumulative gain (NDCG), a widely used measure for evaluating information retrieval results. NDCG has several properties that make it well suited for our setting [Järvelin and Kekäläinen 2002]: (1) NDCG is able to handle nonbinary relevance assessments, (2) it discounts results at lower ranks, and (3) it allows penalizing cases where a method fails to provide a recommendation.

Let r_j denote the rating that the user assigns to the promotion at rank j. The *discounted cumulative gain* (DCG) at rank j equals:

$$DCG[j] = \begin{cases} r_j & j = 1, \\ DCG[j-1] + r_j & 1 < j < b, \\ DCG[j-1] + r_j / \log_b(j) & j \geq b, \end{cases} \tag{11}$$

where b defines the base of the logarithm that is used to discount results. Smaller logarithm bases cause a sharper discounting. We used the base two logarithm, which means our results are discounted from rank two onward. The results of the different items are added up to form a total score. Let $\widehat{r_j}$ denote the maximum possible value of r_j. In our case, we have $\widehat{r_j} = 5$. By replacing r_j in Equation (11) with $\widehat{r_j}$ for all j, we obtain a so-called *ideal DCG* value. As the name suggests, the ideal DCG measures how well the system could perform in the ideal case. The normalized discounted cumulative gain (NDCG) is defined as the DCG divided by the ideal DCG vector. In our case, the ideal DCG corresponds to the vector (5, 10, 13.15, 15.65, 17.81).

Results. Table I presents summary statistics about the ratings provided by the participants. The results indicate that all methods can identify promotions that are considered both relevant and interesting. In terms of relevance, PR and PR-CPE perform better than the baseline on the top three ranks. However, from rank four onward, the baseline and PR perform equally well and obtain slightly better overall ratings than PR-CPE. In terms of interestingness, PR-CPE outperforms the other two approaches, whereas PR performs equally well as the baseline from rank two onward. However, a more careful analysis on the differences between PR-CPE and the other two approaches reveals that PR-CPE is always able to return a promotion, whereas the PR and baseline



(a) Relevance

(b) Interestingness

Fig. 3. Comparison of the NDCG values (y-axis) of the different methods at different ranks (x-axis).

methods often fail to identify any relevant promotions at the lower ranks. Thus, the results indicate that the CPE phase improves recall as it helps to identify a larger set of promotions that are potentially interesting to the customer.

To further explore the tradeoff between the relevance of promotions and the number of promotions that are generated, Figure 3 illustrates the NDCG values of the different methods. From the figure we can observe that the performance of the PR-CPE is better compared to the other methods particularly at the lower ranks, whereas the performance of PR and baseline drop at lower ranks due to the fact that they fail to identify any promotions. Without CPE, the performance of PromotionRank decreases to the same level as the performance of the baseline. The differences are larger in terms of interestingness than in terms of relevance. To determine the significance of the differences, we used MANOVA to compare the NDCG values of the algorithms. At ranks one and two, no statistically significant differences were found between the methods. From rank three onward, PR-CPE outperformed both PR and the baseline ($\lambda = 0.93$, $F(2{,}167) = 3.70$, $p < 0.05$). The analysis revealed no significant differences between PR and the baseline. To further investigate the significance of the responses, we used Bonferroni-adjusted ANOVA to compare the NDCG values of relevance and interestingness separately. At rank one, we found no statistically significant differences

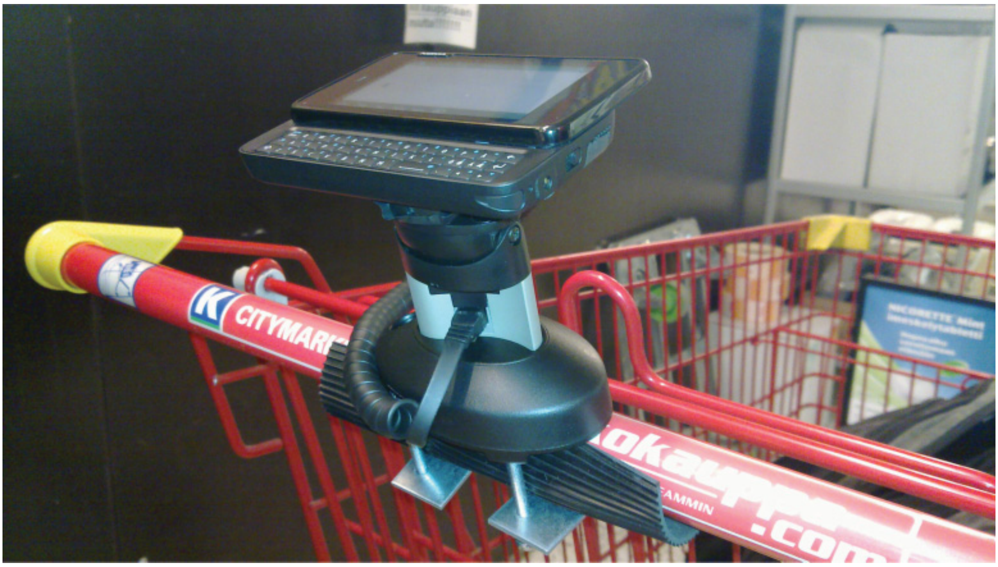


Fig. 4. The system setup used in the field evaluation.

between the different methods. At rank two, the methods performed equally in terms of relevance. In terms of interestingness, the analysis revealed a significant difference between PR-CPE and the baseline ($F(2,167) = 3.31$, $p < 0.05$). The same effect was found on all successive ranks. With regards to relevance, PR-CPE outperformed the baseline from rank four onward ($F(2,167) = 3.45$, $p < 0.05$). The univariate ANOVA revealed no significant differences between PR and PR-CPE. As part of the analysis, we also used ANOVA to verify that the length of the shopping lists had no significant effects on the participants' evaluations of relevance or interestingness. No statistically significant relationships were found with any of the methods, which indicates that our method is robust against variations in the length of the user input.

5.4. Field Evaluation

To ensure PromotionRank is able to identify relevant promotions also in real-world situations, we have conducted a follow-up study within our partner supermarket. This evaluation of PromotionRank was conducted as part of a larger study. In the study, participants were given a grocery aid that they used as part of their shopping visit. The participant's personal shopping list was entered into the grocery aid and the participant was told that the client would present information about currently available promotions during the shopping visit. The grocery aid that we used consisted of a Nokia N900 smartphone, which was attached to the handlebar of a shopping cart using an antitheft connector (see Figure 4). The promotions were presented in the mobile client. The only other functionalities available in the mobile client were the visualization of the participant's shopping list and the possibility to toggle the status of shopping list items as the corresponding products were picked up.

The promotions that were used in the study consisted of the actual special offers for the supermarket where the study was conducted. These offers were automatically collected from the website of the retail chain to which the supermarket belongs. As the supermarket chain has numerous stores, each having a separate set of special offers, we verified before running the study that the products on special offer could actually

Table II. Summary of the Responses to the Questions Measuring the Relevance of the Special Offers Shown in the Field Study
The values in the table correspond to the median and, in parenthesis, the interquartile range of the responses.

		Baseline	PR-CPE
1.	The promoted product fits well with the contents of my shopping list.	3 (2)	4 (3)
2.	The promoted product is a good supplement for the items on my shopping list.	3 (2)	4 (2)
3.	The promoted product is a good substitute for my standard product of choice.	3 (2)	4 (1.5)

be found from the supermarket and that they also had been clearly marked as offers within the supermarket.

At the beginning of the study, each participant was randomly assigned to one of two conditions:

(1) **Random:** The shopping aid presented the participants with a randomly chosen promotion from the set of currently available promotions.
(2) **Personalized:** The shopping aid presented personalized promotions that were determined using the PromotionRank algorithm.

The effectiveness of promotions typically depends on the stage at which they are shown. For example, customers tend to be more susceptible to promotions during the early stages of the shopping visit than later on. The promotion strategies that retail chains use are elaborately designed to take this, and a wide variety of other factors, into account. Consequently, we chose a random baseline for the user study to ensure there was sufficient diversity in the promotions that were presented and to assess whether PromotionRank can provide additional benefits over the promotion strategies that are currently employed by the retail chains.

During the shopping visit, the mobile client triggered promotions at predefined intervals. The first promotion was triggered after 60 seconds, after which the interval was increased by 30-second increments until it reached 150 seconds, which was then used for all subsequent promotions. The promotions were displayed in the mobile client next to the participant's shopping list. An audio alert was played whenever a new promotion became available. To ensure the participants could easily see the advertisements, the mobile phone was programmed to remain in active state (i.e., no screensaver) while the client was running.

After the entry of a participant's shopping list into the shopping client, we let each participant carry out his or her shopping visit without our interference. When the participant arrived at the cashier, we asked the participant to finish the study by answering a set of questions on the promotions that the shopping client had shown. These promotions were replayed in a randomized order. For each promotion, each participant was asked to respond to a three-item questionnaire that measured the relevance of the corresponding promotion. Responses for the questionnaire were elicited using a 5-point Likert scale that was anchored at "completely disagree" (=1) and "completely agree" (=5). We consider a promoted product relevant to the customer if (1) it fits well with the contents of the participant's shopping list, (2) it supplements the contents of the participant's shopping list, or (3) it is a good alternative product to what the customer would otherwise purchase. To measure relevance, we included one item for each of these three possible explanations of relevance; see Table II for the items that were included in the questionnaire. We also collected information about participant demographics and the purchases that the customer made. As part of the analysis, we separately examined correlations between purchase behavior and special offers available at the time of the study.

Participants. Participants for the study were recruited as they entered the store. We recruited individual shoppers as well as families or couples entering the store. The participants were told that they were participating in a usability study of a shopping aid that presents the participant with special offers during their shopping visit. The participants were compensated with a 10€ gift card. To ensure each participant would be exposed to a sufficient amount of promotions, only persons or families with at least four products to purchase were recruited for the study. The sample $N = 22$ consisted of nine females, nine males, and four couples. The random advertisements were shown to 10 participants and the personalized promotions generated were shown to 12 participants. The median age of the participants was 46 years ($IQR = 31$). The length of the shopping lists varied between four and 10 ($MD = 5, IQR = 2$).

Results. The results of the user responses are summarized in Table II. The promotions identified using PromotionRank were considered better than the random promotions in two of the three dimensions of relevance that were considered. Specifically, promotions identified using PromotionRank were considered significantly[5] better in terms of their fit to the contents of the shopping list ($\chi^2 = 7.21$, $p < .05$) and they were considered significantly better supplements than the random promotions ($\chi^2 = 5.92$, $p < .05$). Only the difference as a substitute product was not found significant. We used Poisson regression to verify that the age or gender of the participant had no significant influence on the relevance assessments. We also evaluated whether the length of the shopping list had an influence on the responses. The length of the shopping list had no influence on the responses to the promotions identified using PromotionRank. In contrast, shopping list length had an effect on the random condition where participants with longer lists were more likely to consider the presented promotions relevant ($b = 0.25$, $p < .05$). Together with the results of the first user study, the results of the field study thus demonstrate that PromotionRank can identify relevant and interesting promotions even based on limited natural language inputs. Moreover, the results illustrate that PromotionRank is robust regarding the length of the shopping list that is given as an input to the algorithm.

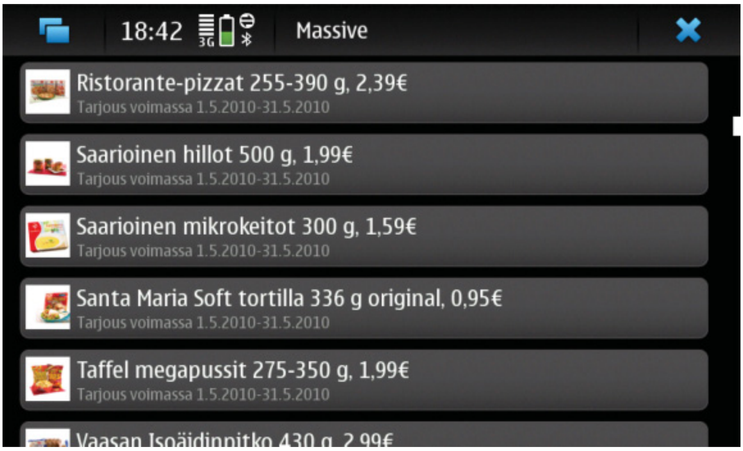
The number of advertisements shown to a participant varied between 2 and 11 (MD = 4, IQR = 1.25). Two of the advertisements shown in the random condition resulted in an additional purchase, whereas four of the advertisements shown in the personalized condition resulted in a purchase.

6. SUMMARY AND DISCUSSION

We presented PromotionRank, a technique for generating personalized rankings of promotions from personal shopping lists. We evaluated PromotionRank using a combination of offline experiments and two user studies. The results of these studies indicate that our method is able to identify promotions that are considered relevant and interesting. Our results also indicate that by using a CPE phase to predict potentially relevant items that are not included in the original input, it is possible to identify additional promotions that are potentially useful and interesting to the customer.

We have integrated PromotionRank as part of Ma$$iv€, an intelligent mobile grocery assistant that provides support for the customer during the entire shopping process [Bhattacharya et al. 2012]. In Ma$$iv€, the shopping list entries are inputted using a customized mobile text entry technique that takes into account correlations between shopping list items [Nurmi et al. 2009b]. Ma$$iv€ uses PromotionRank both to provide

[5] Significance tests were conducted using Kruskal-Wallis nonparametric ANOVA with Bonferroni correction. The reported significance levels have been adjusted to take into account the Bonferroni correction.



Ristorante-pizzat 255-390 g, 2,39€
Tarjous voimassa 1.5.2010-31.5.2010

Saarioinen hillot 500 g, 1,99€
Tarjous voimassa 1.5.2010-31.5.2010

Saarioinen mikrokeitot 300 g, 1,59€
Tarjous voimassa 1.5.2010-31.5.2010

Santa Maria Soft tortilla 336 g original, 0,95€
Tarjous voimassa 1.5.2010-31.5.2010

Taffel megapussit 275-350 g, 1,99€
Tarjous voimassa 1.5.2010-31.5.2010

Vaasan Isoäidinpitko 430 g, 2,99€

Fig. 5. Screenshot of the special offers view in Ma$$iv€ .

a personalized ranking of special offers and to trigger personalized advertisements when the customer is inside a store; see Figure 5 for a screenshot of how the personalized ranking of special offers is displayed in the user interface on a Nokia N900 mobile phone.

While the article has focused exclusively on the retail domain, PromotionRank is applicable to other online domains as well. The combination of user interest modeling and predictions about additional items of interest could be used (e.g., to target advertisements in an online store or as part of an online search engine). In these domains, user interests could be observed (e.g., through web browsing behavior), and the predictions could be based either on personal or aggregate purchase patterns. Another example is contextual advertising, where the user interests could be captured through a combination of search queries and advertisement clickstreams [Chakrabarti et al. 2008; Lacerda et al. 2006]. Moreover, our techniques could also be used to implement an interface that supports browsing through promotions with the help of natural language queries, by assuming that each query constitutes a shopping list with a single item. In this case, the CPE phase would essentially correspond to a query expansion phase.

Previous work in intelligent retail has focused on techniques that require accessing personal shopping history data. We demonstrated that access to customers' shopping history is not a necessary requirement for intelligent retail, and that relevant promotions can be identified, even if the only information about a customer's purchase interests is expressed using natural language. Additionally, since PromotionRank does not rely on historical information, it can provide adequate support even to first-time customers. PromotionRank can thus complement other techniques that rely on personal shopping history data when this data is available. Our focus was exclusively on identifying promotions that have relevant content or are otherwise interesting to the customers. However, additional factors can be easily incorporated into PromotionRank by modifying the category rank scores (Equation (10)). For example, if recent price information would be available, the difference (or ratio) between the average sale price and the promotion price could be used as a measure of the attractiveness of the promotion to the customer.

Currently, the main limitation of PromotionRank is that the predictions that are made in the CPE phase have somewhat low accuracy. This means that the candidate pool is expanded with many nonrelevant categories and that the added categories do not necessarily match well with the shopping list of the customer. This decreases the rank of the promotions that result from these categories. The overall accuracy of the CPE phase depends on the overall number of product categories and the sparsity of data. The CPE phase is a step that would most highly benefit from access to historical shopping list data. Nonetheless, our results also demonstrate that despite the challenges in providing accurate product category predictions, overall, the use of the CPE phase improves the resulting promotion recommendations.

ACKNOWLEDGMENTS

The authors are grateful to their present and past colleagues in the project.

REFERENCES

G. Adomavicius and A. Tuzhilin. 2001. Using data mining methods to build customer profiles. *IEEE Computer* 34, 2, 74–82.

E. Airio. 2006. Word normalization and decompounding in mono- and bilingual IR. *Information Retrieval* 9, 3, 249–271.

P. Barwise and C. Strong. 2002. Permission-based mobile advertising. *Journal of Interactive Marketing* 16, 1, 14–24.

A. Bellogin, P. Castells, and I. Cantador. 2011. Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the 2011 ACM Conference on Recommender Systems (RecSys)*. ACM, 333–336.

S. Bhattacharya, P. Floréen, A. Forsblom, S. Hemminki, P. Myllymäki, P. Nurmi, T. Pulkkinen, and A. Salovaara. 2012. Ma$$iv€ - an intelligent shopping assistant. In *Proceedings of the 8th International Conference on Intelligent Environments (IE)*. IEEE, 165–172.

R. C. Blattberg, R. Briesch, and E. J. Fox. 1995. How promotions work. *Marketing Science* 14, 3, 122–132.

T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. 1999. Using association rules for product assortment decisions: a case study. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 254–260.

D. Chakrabarti, D. Agarwal, and V. Josifovski. 2008. Contextual advertising by combining relevance with click feedback. In *Proceedings of the 17th International Conference on World Wide Web*. ACM, 417–426.

P. Cremonesi, F. Garzotto, and R. Turrin. 2012. Investigating the persuasion potential of recommender systems from a quality perspective: An emperical study. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 2, 11–41.

P. Cremonesi, Y. Koren, and R. Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 2010 ACM Conference on Recommender systems (RecSys)*. ACM, 39–46.

M. Crochemore and W. Rytter. 2002. *Jewels of Stringology*. World Scientific Publishing.

C. Cumby, A. Fano, R. Ghani, and M. Krema. 2004. Predicting customer shopping lists from point-of-sale purchase data. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. ACM, 402–409.

M. Dahlén, A. Rasch, and S. Rosengren. 2003. Love at first site? a study of website advertising effectiveness. *Journal of Advertising Research* 43, 25–33.

A. Demiriz. 2004. Enhancing product recommender systems on sparse binary data. *Data Mining and Knowledge Discovery* 9, 2, 147–170.

M. Deshpande and G. Karypis. 2004. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems (TOIS)* 22, 1, 143–177.

V. Hollink, J. Kamps, C. Monz, and M. de Rijke. 2004. Monolingual document retrieval for European languages. *Information Retrieval* 7, 1–2, 33–52.

M. Hupfer and A. Grey. 2005. Getting something for nothing: The impact of a sample offer and user mode on banner ad response. *Journal of Interactive Advertising* 6, 1.

K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4, 422–446.

B. Keaggy. 2011. *Milk Eggs Vodka: Grocery Lists Lost and Found*. How Books.

J. K. Kim, Y. H. Cho, W. J. Kim, J. R. Kim, and J. H. Suh. 2002. A personalized recommendation procedure for internet shopping support. *Electronic Commerce Research and Applications* 1, 301–313.

Y. Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 426–43.

V. Kumar and R. P. Leone. 1988. Measuring the effect of retail store promotions on brand and store substitution. *Journal of Marketing Research (JMR)* 25, 2, 178–185.

A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto. 2006. Learning to advertise. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 549–556.

R. Lal and C. Matutes. 1994. Retail pricing and advertising strategies. *Journal of Business* 67, 3, 345–370.

R. D. Lawrence, G. S. Almasi, V. Kotlyar, M. S. Viveros, and S. S. Duri. 2001. Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery* 5, 11–32.

G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* 7, 1, 76–80.

G. L. Lohse and P. Spiller. 1998. Electronic shopping. *Communications of the ACM (CACM)* 41, 7, 81–88.

C. Mauri. 2003. Card Loyalty. A new emerging issue in grocery retailing. *Journal of Retailing and Consumer Services* 10, 13–25.

A. Mild and T. Reutterer. 2001. Collaborative filtering methods for binary market basket data analysis. In *Proceedings of the 6th International Conference on Active Media Technology (AMT)*, J. Liu, P. C. Yuen, C. H. Li, J. K.-Y. Ng, and T. Ishida, Eds. Lecture Notes in Computer Science Series, vol. 2252. Springer, 302–313.

C. Monz and M. de Rijke. 2002. Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian. In *Revised Papers from the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*. Lecture Notes in Computer Science Series, vol. 2406. Springer-Verlag, 1519–1541.

P. Nurmi, A. Forsblom, and P. Floréen. 2009a. Grocery product recommendations from natural language inputs. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization*. LNCS Series, vol. 5535. Springer, 235–246.

P. Nurmi, A. Forsblom, P. Floreen, P. Peltonen, and P. Saarikko. 2009b. Predictive text input in a mobile shopping assistant: Methods and interface design. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI)*. ACM, 235–246.

P. Nurmi, E. Lagerspetz, W. Buntine, P. Floréen, and J. Kukkonen. 2008a. Product retrieval for grocery stores. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 781–782.

P. Nurmi, E. Lagerspetz, W. Buntine, P. Floréen, J. Kukkonen, and P. Peltonen. 2008b. Natural language retrieval of grocery products. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. ACM, 1413–1414.

R. Rettie, U. Grandcolas, and B. Deakins. 2005. Text message advertising: Response rates and branding effects. *Journal of Targeting, Measurement and Analysis for Marketing* 13, 4, 304–312.

S. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. 1995. Okapi at TREC-4. In *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*. 73–96.

S. Robertson, H. Zaragoza, and M. Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 42–49.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*. ACM Press, 232–241.

K. Sparck Jones, S. Walker, and S. E. Robertson. 2000a. A probabilistic model of information retrieval: development and comparative experiments Part 1. *Information Processing and Management* 36, 779–808.

K. Sparck Jones, S. Walker, and S. E. Robertson. 2000b. A probabilistic model of information retrieval: development and comparative experiments Part 2. *Information Processing and Management* 36, 809–840.

A. Thomas and R. Garland. 1996. Susceptibility to goods on promotion in supermarkets. *Journal of Retailing and Consumer Services* 3, 4, 233–239.

A. Thomas and R. Garland. 2004. Grocery shopping: List and non-list usage. *Marketing Intelligence & Planning* 22, 623–635.

M. M. Tsang, S.-C. Ho, and T.-P. Liang. 2004. Consumer attitudes toward mobile advertising: An empirical study. *International Journal of Electronic Commerce* 8, 3, 65–78.

D. Vakratsas and T. Ambler. 1999. How advertising works: What do we really know? *Journal of Marketing* 63, 26–43.

M. Vanhuele and X. Drèze. 2002. Measuring the price knowledge shoppers bring to the store. *Journal of Marketing* 66, 72–85.

B. Vindevogel, D. V. der Poel, and G. Wets. 2005. Why promotion strategies based on market basket analysis do not work. *Expert Systems with Applications* 28, 583–590.

R. G. Walters. 1991. Assessing the impact of retail price promotions on product substitution, complementary purchase and interstore sales displacement. *Journal of Marketing* 55, 2, 17–28.

C. Wang, P. Zhang, R. Choi, and M. D'Eredita. 2002. Understanding consumer attitude towards advertising. In *Proceedings of the 8th Americas Conference on Information Systems (AMCIS)*. 1143–1148.

Received July 2012; revised March 2013; accepted May 2013