IRIT at CLEF 2015: A product search model for head queries

Lamjed Ben Jabeur, Laure Soulier, and Lynda Tamine

Paul Sabatier University - IRIT,
Toulouse, France
{jabeur,soulier,tamine}irit.fr

Abstract. We describe in this paper our participation in the product search task of LL4IR CLEF 2015 Lab. This task aims to evaluate, with living labs protective point of view, the retrieval effectiveness over e-commerce search engines. During the online shopping process, users would search for interesting products and quickly access those that fit with their needs among a long tail of similar or closely related products. Our contribution addresses *head* queries that are frequently submitted on e-commerce Web sites. Head queries usually target featured products with several variations, accessories, and complementary products. We propose a probabilistic model for product search based on the intuition that descriptive fields and the category might fit with the query. Finaly, we present results obtained during the second round of the product search task.

Keywords: Information retrieval, product search, e-commerce, BM25F, living labs

1 Introduction

In the last few years, online retailers and marketplaces have shown steady growth in terms of popularity as well as benefits. Amazon claims more than 240 million products available for sale on the US store amazon.com[1]. The marketplace leader claims by the end of 2014 more than 2 billion items sold worldwide [2]. As the result of the huge quantity of available products, users are facing difficulty to make their choice. The diversity of products in types and characteristics complicates the shopping experience of customers on e-commerce Web sites.

To tackle this problem, online retailers include more and more product search tools as a part of their Web sites. Product search is becoming more important as the search space has grown [13], leading to propose adapted retrieval tools in order to help customers to find their products of interest [4]. One example of product search tool is proposed by Google Shopping for which customers have found the utility with 100 billions of submitted search queries by month [3] .

[1] http://www.ecommercebytes.com/cab/abn/y14/m07/i15/s04

[3] http://www.godatafeed.com/resources/google-shopping-campaigns

In the literature, product search has been addressed as an information retrieval (IR) task bridging e-commerce data and customer's information need formulated during the online shopping process. Previous works have proposed to integrate several features which might be split into two categories. On one hand, the authors mainly focus on the product fields, namely its category and its description. Chen et al. [3] proposed to diversify product search results and to return, among the large collection of similar products, only those significantly different from each other. Product categories and attribute values are used, in this case, to diversify the list of products. Vandic et al. [13] address the different hierarchical classification in online stores and the multiple vocabulary terms used to describe the same product. Based on semantic ontologies, they propose to match similar products and classify them into a universal product category taxonomy. On the other hand, users' preferences and search intent are emphasized leading to a user-centered search process. For instance, Duan et al. [8] address specific customer's need who may look for "cheap gaming laptop" or requires a technical feature. Products are ranked using a probabilistic model that models relevance at the level of attribute preferences.

In this paper, we report our participation to the Living Labs for IR (LL4IR) [12] of CLEF 2015 [9]. We propose a probabilistic model for product search that addresses the problem of head queries on e-commerce Web sites. According to Living labs definition [1, 12], head queries represent the set of most frequent queries on featured products. This type of queries may target featured products. The latter may have several variations, accessories, and complementary products. *"Hello Kitty"* and *"Angry birds"* are two examples of LL4IR queries that are frequently submitted to product search engines of an online toys store. These queries return a variety of products that belong to different categories including dolls, miniatures, puzzles, cards, etc. Similarly to the first category of previous work [3, 13], our model relies on product fields, namely the description and the category. Our probabilistic model for product search ranks products with respect to *a*) the likelihood that product's descriptive fields match the query and *b*) the likelihood that the product's category is relevant with regard of the query.

So far, we evaluate our model using the living labs evaluation paradigm for information retrieval introduced by [1, 12]. We report in this paper results obtained fduring the second round of the testing period.

The remainder of this paper is organized as follows. Section 2 introduces our probabilistic model for product retrieval. We describe in section 3 the experimental setup and the results. Section 4 concludes the paper and presents perspectives.

2 Product search model

In this section, we describe our product search model relying on a probabilistic-based retrieval framework. Our goal is to rank products with regard to user's query q by identifying those both belonging to the most likely category and fitting the user's information need.

Products are commonly described in e-commerce Web site with multiple fields[4]. These fields enable to identify the product (i.e., sku, gtin13, ISBN), describe its purpose (i.e., name, brand, description), list elementary and technical features (i.e., model, speed, weight, color) as well as organizing product collection into a structured hierarchy (i.e., category). For convenience, we assume that a product could be seen as a document and, therefore, we consider in what follows the retrieval model as a document ranking one.

With this in mind and inspired by work of Craswell et al. [5] and Dakka et al. [7], we propose to split a document d into a set of textual elements consisting in its category c_d and its set of descriptive fields $\mathcal{D}_d$. The relevance $p(d|q)$ of document d with respect to query q could be rewritten as $p(c_d, \mathcal{D}_d|q)$ (Equation 1). According to probability rules (Equation 2) and assuming that the document category and description are independent (Equation 3), we obtain the following model:

$$\begin{aligned} p(d|q) &= p(c_d, \mathcal{D}_d|q) && (1) \\ &= p(c_d|q) \cdot p(\mathcal{D}_d|c_d, q) && (2) \\ &\propto p(c_d|q) \cdot p(\mathcal{D}_d|q) && (3) \end{aligned}$$

where $p(c_d|q)$ and $p(\mathcal{D}_d|q)$ express respectively the relevance of category c_d of document d and the topical relevance of document description $\mathcal{D}_d$ with respect to query q. We detail these probabilities in what follows.

2.1 Topical relevance of document description $\mathcal{D}_d$

The topical relevance focuses on the document descriptive field set $\mathcal{D}_d$ and estimates its similarity with the query terms. Except the category field, all the remaining fields are part of the document description $\mathcal{D}_d$. These fields may be less or more effective for product search. Some fields such as the title are usually size limited; so they include concise information about the product. Other may include broader information such as the description field. Fields that report technical features are helpful for technical constrained information need. Yet, head queries in which we interest in this paper do not include such technical constraints but address the overall aspects of the product.

In this aim, we propose to use the $BM25F$ scoring schema [14, 6] to estimate the likelihood $p(\mathcal{D}_d|q)$ that the document descriptive fields $\mathcal{D}_d$ match the query q. The $BM25F$ computes the similarity of document d with respect to query q while giving different importance scores to each field.

First, we calculate a normalized term frequency $\overline{tf}_{t,f,d}$ for each field.

$$\overline{tf}_{t,f,d} = \frac{tf_{t,f,d}}{1 + b_f(\frac{l_{f,d}}{l_f} - 1)} \tag{4}$$

Where $tf_{t,f,d}$ represents the frequency of term t, in the field f belonging to description $\mathcal{D}_d$ of document d. $l_{f,d}$ is the length of field f in document description

[4] http://www.schema.org/product

$\mathcal{D}_d$ and l_f is the average length of the field f. b_f is a field-dependant parameter similar to the b parameter in BM25 [11].

The term frequencies estimated over all the field set are combined linearly using the field weights w_f as follows:

$$\overline{tf}_{t,d} = \sum_{f \in \mathcal{D}_d} w_f * \overline{tf}_{t,f,d} \tag{5}$$

The term frequency $\overline{tf}_{t,d}$ is then integrated in the usual BM25 saturating function [11] that models the non-linear relevance distribution of term frequencies. The similarity of document description $\mathcal{D}_d$ with respect to query q is computed as next:

$$BM25F(q, \mathcal{D}_d) = \sum_{t \in q \cap \mathcal{D}_d} \frac{\overline{tf}_{t,d}}{k_1 + \overline{tf}_{t,d}} idf(t) \tag{6}$$

where k_1 and $idf(t)$ express respectively the BM25 parameter and the inverse document frequency of term t, similarly to [11].

The probability $p(\mathcal{D}_d|q)$ introduced in Equation 3 is approximated by the BM25F function:

$$p(\mathcal{D}_d|q) \approx BM25F(q, \mathcal{D}_d) \tag{7}$$

2.2 The relevance of the category

The relevance of category c_d with respect to query q aims at identifying to what extent the category is relevant over the document collection. The idea behind is to decide which eminent category likely matches the query since different categories may respond to the query.

Let S be the set of non-negative topical scores obtained by document description $\mathcal{D}_d$ of all documents $d \in D(c_d)$, where $D(c)$ correspond to the set of documents characterized by category c_d. More formally, S is defined as follows:

$$S = \{p(\mathcal{D}_d|q) | d \in D(c_d) \wedge p(\mathcal{D}_d|q) > 0\} \tag{8}$$

where $p(\mathcal{D}_d|q)$ is approximated by $BM25F(q, d)$ score as presented in Equations 6 and 7.

We propose to estimate the similarity $sim(q, c_d)$ of document category c_d with regard to the query q as the product of the log scale cardinality of set S and an aggregate function $A(S)$ of topical scores over respective documents:

$$sim(q, c_d) = \log(1 + |S|) * \mathcal{A}(S) \tag{9}$$

where $\mathcal{A}(S)$ can be computed as the maximum, the mean and the median scores over the topical distribution of all documents $D(c_d)$. We propose to use the 95^{th} percentile as aggregate function $A(S)$. In contrast of mean and maximum, the 95^{th} percentile is resistant. Similarly to the median, 95^{th} percentile allows to measure the global tendency of topical scores.

As the category includes more relevant documents with respect to the query, the category might be relevant to the query. This is reflected by the first part of Equation 9, noted $\log(1 + |S|)$. The log scale value enables to attenuate high cardinality and thus corrects the importance of overpopulated categories.

In connection to equation 3, we note that the likelihood $p(c_d|q)$ that the document category c_d is relevant in regard of query q is approximated to the similarity $sim(q, c_d)$ of document category c_d with regards to the query q.

$$p(c_d|q) \approx sim(q, c_d) \tag{10}$$

3 Experimental evaluation

Living Labs for IR (LL4IR) [12, 2] provides a benchmarking platform for evaluating information retrieval effectiveness. The benchmarking platform is implemented as a cloud service. The performances of our participating system are evaluated with real users in real environments. As long as users generate feedback about displayed ranking, a comparison to the production system is immediately available. We note that participant system must be computed offline and submitted to benchmarking service though a REST API.

Two search scenarios are evaluated this year including product search and Web search. We participated only to the product search scenario. In order to build our runs, we follow the next steps:

1. We gather the query set from query API resource *"participant/query"*
2. We get for each query the list of candidate documents to be ranked. This list is available through the doclist API resource *"participant/doclist"*.
3. For each document ID in the list, we get the respective document content via the document API resource *"participant/doc"*.
4. We apply Snowball stemming algorithm on document textual fields. We used Hungarian Snowball stemmer provided by Lucene Java Library.
5. We compute document scores as presented in equation 3 then we rerank document by descending score order.
6. We format run and submit it to run API resource *"participant/run"*.

3.1 Parameter Setup

Table 1 lists available fields for product search and the respective weights used in our run. These weights highlight important fields or discard irrelevant ones. We note that weights are used to compute the BM25F score presented in Equation 5. Compared to the title field in BM25F empirical experiments [14], we propose also to give a highest weight to the name field ($product\_name = 38.4$). We give higher weights (35) to brand and characters, which are comparable to anchor field in [14]. As head queriers often include the brand name or named entities that correspond in this use case to characters. We also consider category name and short description which are similar to body field with minimal weights with a value equals to 1. The remaining fields are discarded with respective weights equals to 0 including the description field. In fact, the description field in the case of e-commerce may include technical features or boarder information that is not helpful in the case of head queries. Furthermore, the description may include a list of compatible assets or complementary product which may assimilated to a term frequency representation. We also note that about 57% of products in LL4IR dataset have an empty description.

Field	Weight	Field	Weight	Field	Weight
age_max	0	age_min	0	arrived	0
available	0	bonus_price	0	brand	35
category	1	category_id	0	characters	35
description	0	main_category	0	main_category_id	0
gender	0	photos	0	price	0
product_name	38.4	queries	0	short_description	1

Table 1. Descriptive field weights

Based on the empirical evaluation of BM25F [14, 6], we set the value of k_1 to 2.0. As most of considered fields are short (i.e. title) or extremely short (i.e. brand), we propose to ignore field length normalization in Equation 4. In accordance, l_f is set to 0 ($l_f = 0$).

3.2 Metrics

According to Living Labs approach, document ranking of participating system is mixed with the document ranking of the production system. The latter corresponds to the default document ranking system provided by Web site owners. For each submitted query belonging to the pre-selected head query set, the user get a set of results for which the half comes from website production system and the other half from a random participating system. Beside comparison to the production system, organizers have implemented a baseline system which is submitted with same conditions as participating system. We note that the baseline is different than the production system. It ranks products based on historical click-through rate [12].

With respect to the approach, 5 metrics are proposed by Living Labs organizers in order to evaluate participating system. These metrics, estimated over all submitted head queries, are presented in what follows:

- The number of wins, noted $\#Wins$, which expresses the number of times the test system received respectively more clicks than the product system.
- The number of losses, noted $\#Losses$, which expresses the number of times the test system received respectively fewer clicks than the product systems.
- The number of ties, noted $\#Ties$, which expresses the number of times the test system received respectively as many clicks as the product systems.

- The number of Impressions, noted $\#Impressions$, which expresses the test system is mixed with production one.

$$\#Impressions = \#Wins + \#Losses + \#Ties \tag{11}$$

- The outcome, noted $Outcome$, is defined as the ratio of wins over the sum of wins and losses (Equation 12). A ratio higher than 0.5 highlights the system ability to provide more relevant documents than irrelevant ones, assuming that clicks are indicators of document relevance [10].

$$Outcome = \frac{\#Wins}{\#Wins + \#Losses} \tag{12}$$

3.3 Results

Table 2 presents the results obtained for our model on the testing query set with respect to the baseline performances and the different participants of the second round from Jun 15, 2015 till Jun 30, 2015.

Run	Outcome	#Wins	#Losses	#Ties	#Impressions
Baseline	0.5284	93	83	598	774
UiS-Jern	0.4795	82	89	596	767
GESIS	0.4520	80	97	639	816
UiS-Mira	0.4389	79	101	577	757
UiS-UiS	0.4118	84	120	527	731
IRIT	0.3990	79	119	593	791

Table 2. Results of the second testing period

From a general point of view, we highlight that the baseline overpassed the whole set of participants and reached an outcome value higher than 0.5 while participant systems obtained outcome values lower than 0.5. This latter result shows that the number of losses for both system exceeds the number of wins for all participants systems.

Concerning our model, we obtained the lowest outcome value equals to 0.399. Since we do not have information about other participant system, we are not able to explain this result. In contrast to our model only based on field content and distribution over the collection, the baseline uses click-through rate to generate a historical ranking. The use of relevance feedback of user feature, often used in classical information retrieval to rank documents [10], explains the important differences between the obtained results.

Since our model provides low results, we propose here a query analysis investigating how much our model is effective at the query level. Our objective is to identify categories of queries for which our model is effective. Figure 1 illustrates the results obtained query by query. Over the 50 queries, we obtained an evaluation metric higher than 0.5 for 20 queries. One can notice that queries related to the most famous brands (*"Scrabble"*, *"Fisher Price"*, *"Poni"*, *"Playmobil"*, *"Angry Birds"*, etc) obtained an outcome value higher than 0.5, except queries R-q54 and R-q89 dealing with the *"Lego"* and *"Lego-city"* products. Therefore, we believe that our model is adapted to solve head queries that address popular brands and characters.

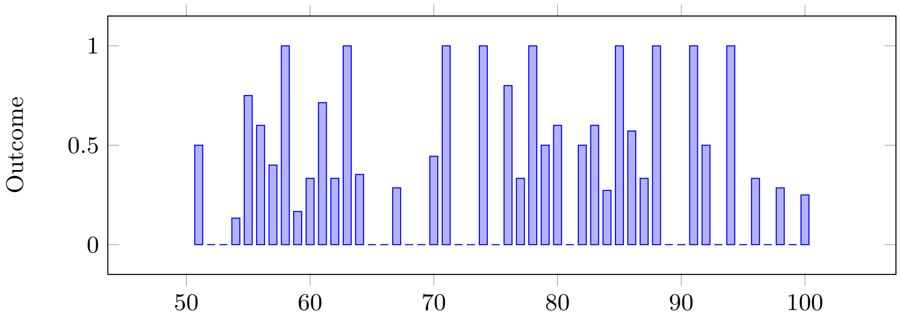

Fig. 1. Effectiveness analysis at the query level

4 Conclusion

In this paper, we presented a product-search-based probabilistic model relying on two types of product features, corresponding either to textual or hierarchical features. This lead us to propose two types of scores based on the topic as well as the category relevance. Experiments of the second round highlight lower results that those obtained for the baselines and by other participants. We also showed that our model is more effective for search related to famous product.

These statements highlight that product search, and more particularly in a Living-Labs setup, is a difficult and novel task in information retrieval which would gain in maturity with more explorations and deeper work in the domain. In terms of perspectives, we believe that further work on the failure analysis highlighting the relevant features adapted to product search by comparing the obtained results with the users' search intent would benefit to better understand this particular search task. Then, we plan to tune our model with relevant features highlighted by this deep analysis.

Bibliography

[1] Krisztian Balog, Liadh Kelly, and Anne Schuth. Head first: Living labs for ad-hoc search evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1815–1818, New York, NY, USA, 2014. ACM.

[2] L. Cappellato, N. Ferro, G. Jones, and editors (2015) San Juan, E. Clef 2015 labs and workshops, notebook papers. In *CEUR Workshop Proceedings (CEUR-WS.org)*, ISSN 1613-0073, 2015.

[3] Xiangru Chen, Haofen Wang, Xinruo Sun, Junfeng Pan, and Yong Yu. Diversifying product search results. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1093–1094, New York, NY, USA, 2011. ACM.

[4] Brian J Corbitt, Theerasak Thanasankit, and Han Yi. Trust and e-commerce: a study of consumer perceptions. *Electronic commerce research and applications*, 2(3):203–215, 2003.

[5] Nick Craswell, Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 416–423. ACM, 2005.

[6] Nick Craswell, Hugo Zaragoza, and Stephen Robertson. Microsoft cambridge at trec 14: Enterprise track.

[7] W. Dakka, L. Gravano, and P.G. Ipeirotis. Answering general time-sensitive queries. *Knowledge and Data Engineering, IEEE Transactions on*, 24:220–235, 2012.

[8] Huizhong Duan, ChengXiang Zhai, Jinxing Cheng, and Abhishek Gattani. Supporting keyword search in product database: A probabilistic approach. *Proc. VLDB Endow.*, 6(14):1786–1797, September 2013.

[9] Mothe J., Savoy J., Kamps J., Pinel-Sauvagnat K., Jones G., San Juan E., Cappellato L., and Ferro N. (ed.). Experimental ir meets multilinguality, multimodality, and interaction. In *Sixth International Conference of the CLEF Association*, CLEF'15. LNCS, vol. 9283, Springer, Heidelberg, 2015.

[10] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, 2002.

[11] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[12] Anne Schuth, Krisztian Balog, and Liadh Kelly. Overview of the living labs for information retrieval evaluation (ll4ir) clef lab 2015. In *CLEF 2015 - 6th Conference and Labs of the Evaluation Forum*, Lecture Notes in Computer Science (LNCS). Springer, 2015.

[13] Damir Vandic, Jan-Willem van Dam, and Flavius Frasincar. Faceted product search powered by the semantic web. *Decision Support Systems*, 53(3):425 – 437, 2012.

[14] Hugo Zaragoza, Nick Craswell, Michael J Taylor, Suchi Saria, and Stephen E Robertson. Microsoft cambridge at trec 13: Web and hard tracks. Citeseer.