Узнать или купить? Классификатор страниц обзоров и интернет-магазинов

To find out or to buy? Product review vs. Web shop classifier

Braslavski P. I. (pb@yandex-team.ru), Yandex

Kiselev Yu. A. (yurikiselev@yandex-team.ru), Ural Federal University

In this paper we examine two categories of search results retrieved in response to product queries. This classification reflects the two main kinds of user intents — product reviews and online shops. We describe the training and test samples, classification features, and the classifier's structure. Our findings demonstrate that this method has good quality and performance suitable for real-world applications.

1. Introduction

Recently the diversity of search results has gained the attention of information retrieval researchers and practitioners. When considering the diversity of search results we shift the emphasis from the relevance of a single query-document pair to the relationships between documents in the result list and search engine results page (SERP) as a whole. The diversity of search results has many aspects [1] that are associated with the incompleteness of available information. Ambiguous queries are a classic example, when a diverse SERP could compensate for the lack of knowledge about the actual needs of the user. For instance, the Russian query [*алые паруса*] (*scarlet sails*) may refer to the novel by Alexander Grin, its screen version of 1961, a retail chain, a residential complex in Moscow, or a school graduation day celebration in St. Petersburg. When it is impossible to disambiguate the query, we can try to organize the result list the way it reflects the different intents of the query (how to identify these intents and what method to choose to structure SERP are other problems). Another aspect is the ambiguity associated with the actual users. For example, the query [*fixed assets amortization*] might be issued by an experienced accountant or a college student doing her/his homework. Accordingly, if we cannot obtain additional information, the search results may contain documents addressing topics on different levels. Another consideration is the *genre* variety of documents in the results: documents on the same topic, but of different types. For example, the results for [*large hadron collider*] may contain both news and popular science articles.

In our work, we consider the problem of diversity for queries about the products traditionally offered in online shops. The spectrum of these products is well presented on the shopping comparison service Yandex.Market (http://market.yandex.ru) and includes electronics, photo and home appliances, mobile phones, computers, etc. The typical examples of online shopping queries are queries like [*samsung g400*], [*home air conditioner*], and [*netbooks review*]. We estimate the share of such queries at about 4% of the whole query stream on Yandex. The range of users' needs behind such queries can be quite broad. However, the majority of users either want to: 1) know what is being offered, make the choice, examine the product's characteristics, compare it with similar products — these are the steps usually leading to the purchase of a product presented on Yandex.Market — or 2) make the actual order or purchase. These user intents correspond to two types of documents: 1) online product surveys and reviews and 2) webshop pages where users can make an order. Of course, these intents do not cover the full range of users' needs — people may use the same queries to search for technical documentation, spare parts, service and repair shops, accessories, software for devices, classified ads, etc. However, the two aforementioned needs are prevalent.

In our work, we are not offering methods to achieve search result diversity, but showing, instead, how to create preconditions for it by addressing the problem of classifying web documents into reviews and online shops (see for example [2] on diversity-based search results optimization). We define as "reviews" detailed and thorough professional or editorial reviews, while excluding short user opinions. Digital Photography Review (http://dpreview.com/) is a good example of such content.

In the rest of the paper we give an overview of related work on web document classification (Section), describe the requirements for and the resulting structure of classifiers (Section), specify our data (Section), define classification features (Section 0), and present evaluation results (Section 6). Section 7 is the conclusion.

2. Related work

Various web page classifications are widely used in web applications, including web search. Web document categorization is used to improve search quality, build vertical searches, filter spam, to categorize user queries, etc. In contrast to traditional methods of text document classification, web page classification can be based on a wider range of features including those based on document structure, HTML tags, metadata, hyperlinks, URLs and user behavior. The problem of classifying web documents can be complicated by clutter such as advertisements, navigation bars, etc. Since the pioneering work by Joachims [4] SVM (Support Vector Machine) is a method widely used to classify text documents.

Page classification into reviews and online shops is an example of genre classification. A detailed survey of the approaches to and methods of genre classification is presented in [5]. At least two noteworthy papers dealing with the analysis of web documents appeared after the survey had been published. Meyer zu Eissen and Stein [6] conducted a user study spawning a set of eight web genres useful for web search,

and built a corpus containing these genres. Along with the surface and linguistic features traditionally used in genre analysis, their study employed HTML-based features. The method was implemented as a plug-in for the Firefox browser that enriches Google snippets with genre labels [7]. Lim et al. [8] expanded this approach even further and made use of a wider range of features (326 in total), including various surface, lexical, syntactic, HTML, and URL features.

Mindset, a Yahoo! research project [9], allowed users to rank search results based on their commercial or informational value. In addition to the standard query box, Mindset had a slider that the user could move between «shopping» and «researching», changing the appearing results from less to more commercial. Unfortunately, the project is now closed, and the implementation details have not been published.

Dai et al. [10] solved the problem of detecting user's online commercial intention. In order to accomplish this task they constructed a classifier of commercial and noncommercial web pages. Classification was performed using SVM in the space of terms, term occurrences in the document's body and HTML-tags were counted separately (thus, n terms generated $2n$ features). The training sample contained 5,375 pages, 2,820 of them were labeled as commercial. The authors obtained good results with precision 93.0 % and recall 92.5 % for the commercial pages class. The demo classifier is available online [11].

Paper [12] describes a simple client-side tool that classifies commercial (i. e. online shops' product pages) vs. noncommercial pages. Classification is performed based on different features: presence of images and product descriptions, indication of price, "buy" button, URL, etc. Classification is followed by product name and price extraction.

The problem of filtering product reviews from search results is addressed in [13]. The task was solved based on result snippets: experimental dataset contained 1,200 Google snippets for queries in the form [*product_name* + *"review"*]. The features used for classification were terms in the title, URL, and snippet itself. The final classifier combined the result of SVM classifier and heuristic rules.

Product review classification, based on label propagation over click graphs was considered, among other classification problems, in [14]. A sample of 10,000 positive and negative examples was used for learning with gradient boosting of decision trees. Different features were used: text (unigrams and bigrams in various structural parts of the document, the number of words in the document, the number of capitalized words), link (properties of incoming and outgoing links), URL (length and presence of certain tokens), and HTML features (presence of specific tags). The best results for review class reported by the authors: precision — 63.96 %, recall — 73.97 %.

3. Classifier

Our goal was to build a classifier suitable for a large-scale web search engine capable to process billions of web pages in reasonable time. So, performance was as crucial as the quality of classification. Consequently we were restricted to employ only light-weight features that could be extracted by one-pass page scan. We opted for embedding the classifier into the search engine's indexing pipeline. Even though it led

to even harder efficiency restrictions, we could easily employ tokenization, lemmatization, language detection and other results available at indexing time.

For learning we used LIBSVM [15], an implementation of SVM. To compose a three-class classifier out of binary classifiers (`shop – other`, `review – other`) we had two options:

- ***Parallel classifier.*** The page is processed by both classifiers independently. As a result, some pages can be assigned to both classes (`shop` and `review`).
- ***Sequential classifier.*** Negative (`other`) output of the `shop` classifier is fed to `review` classifier.

In fact, these options differ insignificantly. In both cases we had to extract all features at once (see Section 5). Since web shop pages account for about 4% (see Section 6) of the web (the reviews share is much less), sequential scheme does not save much computations.

4. Data

To classify a significant portion of the indexed documents (excluding only documents in a language other than Russian and very short documents), we constructed problem-driven training and test sets consisting of the documents returned to product queries on Yandex. This approach supposes that we can automatically detect queries of the target class. The problem of classifying queries is beyond the scope of this paper (for example, [3] describes a method for detecting product queries with high precision and recall).

In order to build the training sample, we randomly sampled 100 queries from the list of manually tagged product queries. For each query we downloaded top10 documents from Yandex SERP. The total number of downloaded pages was 979 (some pages were inaccessible and other were filtered out as non-Russian). The set was labeled by a Yandex assessor. Each web page had to be assigned to exactly one class: `shop`, `review` or `misc`. If a page had properties of both the `shop` and the `review` class (e. g. a shop page with a detailed product description), then it had to be labeled as `shop` (i. e. `shop` label overrides `review` label). Table 1 shows the break-down of the sample.

Table 1. Learning sample for shop classifier

Class	# of pages
Shop	301
Review	87
Misc	591
Total	979

Initial experiments with this sample showed that its size does not allow for a learning review classifier of a satisfactory quality. So, we used this sample for the learning `shop` classifier only. To train the review classifier, we composed

a synthetic learning sample. It contained 150 `review` pages, 150 miscellaneous pages from the initial training sample labeled as `misc`. Also, we added 50 long documents collected manually (biographies, encyclopedia entries, etc.). Table 2 shows this breakdown.

Table 2. Learning sample for review classifier

Class	# of pages
Review	150
Misc	150
Long docs	50
Total	350

The test sample was obtained the same way as the `shop` training sample: we downloaded and labeled top10 from the Yandex results for 100 product queries. Table 3 shows the structure of the test sample.

Table 3. Test sample

Class	# of pages
Shop	431
Review	101
Misc	557
Total	1089

5. Classification features

5.1. Shop classifier

We used different feature groups for classification: term, textual, lexical, HTML, and URL features.

Term features. We identified the most informative term-features based on *mutual information*. For performance reasons, we did not consider the semantic or the visual structure of a document (document's main content, navigation, headers, footers etc.). As expected, the most contrasting terms were *магазин, рубль, каталог, цена, прайс,* and *корзина* (*shop*, *ruble*, *catalog*, *price*, and *basket*). The full list of terms used for classification consisted of about one hundred terms.

HTML features. The main high-level feature of a shop page is a possibility to make an order. We used two features aimed at detecting the "buy" button:

- number of specific keywords (*купить* — *buy, заказать* — *order*, etc.) in links and buttons;

- number of HTML-tags (*img*, *button*, etc.) with words "*cart*", "*basket*", "*order*" etc. in attributes.

Lexical features. We used the list of trademarks and brands on the Yandex.Market comparison shopping service (excluding commonly used words and the names consisting of two and more words). This list generated two features: the number of words from the list on the page and the number of unique words from the list.

URL feature. Many tokens in URLs are good cues for classification of a page as a web shop. This feature reflected the number of specific terms, such as *product*, *shop*, *itemID*, etc. in the URL.

5.2. Review classifier

Term features. By analogy with the shop classifier, we selected the most informative terms for the review classification. Since lexical variety of reviews is much higher than that of shop pages, the list of contrasting words was much longer and exceeded 7,000 words. The most informative terms for review class were *рынок, взгляд, автор, обзор, комментарий, маленький,* and *китайский* (*market*, *view*, *author*, *review*, *comment*, *small*, and *Chinese)*.

Textual features. Textual features were document's length in words and characters and sentence length distribution.

Lexical features. The list of 165 manually collected appraisal adjectives — *хороший, прекрасный, великолепный, плохой, отвратительный, ужасный,* etc. (*good*, *excellent*, *magnificent*, *bad*, *disgusting*, *awful*, etc.) — produced two features: the total number of words from the list and the number of unique words.

6. Results

Classification results with various feature groups for the test sample are presented in tables 4 and 5.

Table 4. Online shop classification results

Set of features	Precision	Recall
Terms only	0.918	0.809
HTML features only	0.894	0.491
Term + HTML features	0.934	0.800
Term + lexical features	0.910	0.807
Term + URL features	0.876	**0.856**
All features	**0.937**	0.837

Table 4 shows that classification based on terms only produced good results. Adding HTML markup features, i. e. detecting "buy" button, increased precision

of classification. These findings support the results shown by Dai et al. [10]: term features and HTML tags work well even with a learning sample of a modest size. The observation that lexical features generated from the list of vendors and brands impair the quality of classification can be explained by the fact that almost all pages returned in response to a commercial query already contain brand names. The features would probably have increased the quality, if we evaluated classification results on a sample of random web pages. Adding URL features reduced precision, but increased recall. A set of all presented features provided the best precision for `shop` class (0.937).

Table 5. Review classification results

Set of features	Precision	Recall
Terms only	0.644	0.861
Term + URL features	0.643	0.841
Term + lexical features	0.625	0.861
Term + textual features	**0.681**	**0.891**

As expected, the quality of `review` classifier was much lower, considering the diversity of the class members and the shallow features we used. The lexical and URL features did not contribute to classification quality. The term and textual features provided the best precision for `review` class (0.681).

Tables 4 and 5 show the results of parallel classification (i.e. the entire test sample was processed by both classifiers, see Section). Superimposition of classifiers' results showed that only 16 pages were assigned both to `shop` and `review` (all these 16 pages were labeled as `shop` by a human assessor). The results of the three-class classifier (`shop` label overrides `review` label) are shown in Table 6 (true classes in rows, classification output in columns).

Table 6. Confusion matrix of the three-class classifier

	Shop	Review	Misc	Recall
Shop	361	3	67	**0.84**
Review	1	90	10	**0.89**
Misc	23	23	511	
Precision	**0.94**	**0.78**		

To check the hypothesis that the `shop` classifier will perform well even on arbitrary documents (not only on documents returned to specific queries), we sampled randomly 56,768 Russian pages from Yandex's index. 2,071 pages were automatically labeled as `shop`, 1,908 of the labels (3.6% of the initial sample) were approved by a human, which resulted in precision 0.92.

7. Conclusion and future work

In this paper we presented a genre classifier classifying search results retrieved by product queries into two classes reflecting the two main intents of the user — product reviews and online shops. The aim of this classification is to compensate for the lack of knowledge about the actual needs of the user by providing a diversity of search results.

In the future our work will center around:

- information extraction from web shop and product review pages: product name, its category, price, etc.;
- improving quality of the product review classification. To bootstrap the results, we plan to calculate linguistically richer features in off-line mode;
- investigating the possibilities on taking page segmentation into account (i. e. page main content, navigation, etc.) to improve classification accuracy, as some studies on web page classification suggest.

References

1. *Filip Radlinski, Paul N. Bennett, Ben Carterette, and Thorsten Joachims.* Redundancy, diversity and interdependent document relevance // SIGIR Forum 43, 2 (December 2009), 46–52.
2. *Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong.* Diversifying search results // WSDM '09, 2009, 5–14.
3. *Xiao Li, Ye-Yi Wang, and Alex Acero.* Learning query intent from regularized click graphs // SIGIR '08, 2008, 339–346.
4. *Thorsten Joachims.* Text Categorization with Support Vector Machines: Learning with Many Relevant Features // ECML-98, 1998, 137–142.
5. *Santini M.* State-of-the-art on automatic genre identification // Technical Report ITRI-04–03, 2004, ITRI, University of Brighton (UK). Available online: ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf
6. *Sven Meyer zu Eissen, Benno Stein.* Genre Classification of Web Pages // KI 2004, 256–269.
7. *WEGA* (Web Genre Analysis) project, http://www.webis.de/research/projects/wega
8. *Lim, C.S., K. J. Lee, and G. C. Kim.* Multiple sets of features for automatic genre classification of web documents // Information Processing & Management, vol. 41, 2005, 1263–1276.
9. *MindSet,* http://research.yahoo.com/node/1912
10. *Honghua (Kathy) Dai, Lingzhi Zhao, Zaiqing Nie, Ji-Rong Wen, Lee Wang, Ying Li.* Detecting online commercial intention (OCI) // WWW'06, 2006, 829–837.
11. *Detecting* Online Commercial Intention http://adlab.msn.com/Online-Commercial-Intention/Default.aspx
12. *Renan Cattelan, Darko Kirovski, Deepak Vijaywargi.* Serving Comparative Shopping Links Non-invasively // Proceedings of the Web Intelligence and Intelligent Agent Technologies, 2009, 498–507.

13. *Tun Thura Thet, Jin-Cheon Na, and Christopher S. G.* Khoo Automatic Classification of Web search results: product review vs. non-review documents // ICADL'2007, 2007, 65–74.
14. *Soo-Min Kim, Patrick Pantel, Lei Duan, and Scott Gaffney.* 2009. Improving web page classification by label-propagation over click graphs // CIKM '09, 2009, 1077–1086.
15. *LIBSVM,* http://www.csie.ntu.edu.tw/~cjlin/libsvm/