Query classification using topic models and support vector machine

Dieu-Thu Le
University of Trento, Italy
dieuthu.le@disi.unitn.it

Raffaella Bernardi
University of Trento, Italy
bernardi@disi.unitn.it

Abstract

This paper describes a query classification system for a specialized domain. We take as a case study queries asked to a search engine of an art, cultural and history library and classify them against the library cataloguing categories. We show how click-through links, i.e., the links that a user clicks after submitting a query, can be exploited for extracting information useful to enrich the query as well as for creating the training set for a machine learning based classifier. Moreover, we show how Topic Model can be exploited to further enrich the query with hidden topics induced from the library meta-data. The experimental evaluations show that this system considerably outperforms a matching and ranking classification approach, where queries (and categories) were also enriched with similar information.

1 Introduction

Query classification (QC) is the task of automatically labeling user queries into a given target taxonomy. Providing query classification can help the information providers understand users' needs based on the categories that the users are searching for. The main challenges of this task come from the nature of user queries, which are usually very short and ambiguous. Since queries contain only several to a dozen words, a QC system often requires either a rather large training set or an enrichment of queries with other information (Shen et al., 2006a), (Broder et al., 2007).

This study will focus on QC in art, culture and history domain, using the Bridgeman art library[1], although our framework is general enough to be used in different domains. Manually creating a training set of queries to build a classifier in a specific domain is very time-consuming. In this study, we will describe our method of automatically creating a training set based on the click-through links and how we build an SVM (Support Vector Machine) classifier with the integration of enriched information. In (Le et al., 2011), it has been shown that click-through information and topic models are useful for query enrichment when the ultimate goal is query classification. We will follow this enrichment step, but integrate this information into a SVM classifier instead of using matching and ranking between queries and categories as in (Le et al., 2011).

The purpose of this paper is to determine (1) whether the query enrichment with click-though information and hidden topics is useful for a machine learning query classification system using SVM; and (2) whether integrating this enriched information into a machine learning classifier can perform better than the matching and ranking system.

In the next section, we will briefly review the main streams of related work in QC. In section 3, we will describe the Bridgeman art library. Section 4 accounts for our proposed query classification framework. In section 5, we will present our experiment and evaluation. Section 6 concludes by discussing our main achievements and proposing future work.

2 Related work

Initial studies in QC classify queries into several different types based on the information needed by the user. (Broder, 2002) considered three different types of queries: informational queries, navigational queries and transactional queries. This stream of study focuses on the type of the queries, rather than topical classification of the queries.

Another stream of work deals with the problem

[1] http://www.bridgemanart.com/

Proceedings of the 2012 Student Research Workshop, pages 19–24,

Jeju, Republic of Korea, 8-14 July 2012.

of classifying queries into a more complex taxonomy containing different topics. Our study falls into this second stream. To classify queries considering their meaning, some work considered only information available in queries (e.g., (Beitzel et al., 2005) only used terms in queries). Some other work has attempted to enrich queries with information from external online dataset, e.g., web pages (Shen et al., 2006a; Broder et al., 2007) and web directories (Shen et al., 2006b). Our work is similar to their in the idea of exploiting additional dataset. However, instead of using search engines as a way of collecting relevant documents, we use the metadata of the library itself as a reference set. Furthermore, we employ topic models to analyze topics for queries, rather than enriching queries with words selected from those webpages directly as in (Shen et al., 2006a; Broder et al., 2007).

The context of a given query can provide useful information to determine its categories. Previous studies have confirmed the importance of search context in QC. (Cao et al., 2009) considered the context to be both previous queries within the same session and pages of the clicked urls. In our approach, we will also consider click through information to enrich the queries and analyze topics.

In (Le et al., 2011), queries and categories are enriched with both information mined from the click-through links as well as topics derived from a topic model estimated from the library metadata. Subsequently, the queries are mapped to the categories based on their cosine similarity. Our proposed approach differs from (Le et al., 2011) in three respects: (1) we enrich the queries, but not the categories (2) we employ a machine learning system and integrate this enriched information as features to learn an SVM classifier (3) we assume that the category of a query is closely related to the category of the corresponding click-through link, hence we automatically create a training data for the SVM classifier by analyzing the query log.

3 Bridgeman Art Library

Bridgeman Art Library (BAL)[2] is one of the world's top image libraries for art, culture and history. It contains images from over 8,000 collections and more than 29,000 artists, providing a central source of fine art for image users.

Works of art in the library have been annotated with titles and keywords. Some of them are categorized into a two-level taxonomy, a more fine-grained classification of the Bridgeman browse menu. In our study, we do not use the image itself but only the information associated with it, i.e., the title, keywords and categories. We will take the 55 top-level categories from this taxonomy, which have been organized by a domain expert, as our target taxonomy.

4 Building QC using topic models and SVM

Following (Le et al., 2011), we enrich queries both with the information mined from the library via click-through links and the information collected from the library metadata via topic modeling. To perform the query enrichment with topics derived from the library metadata, there are several important steps:

• Collecting and organizing the library metadata as a reference set: the library metadata contains the information about artworks that have been annotated by experts. To take advantage of this information automatically, we collected all annotated artworks and organized them by their given categories.

• Estimating a topic model for this reference set: This step is performed using hidden topic analysis models. In this framework, we choose to use latent dirichlet allocation, LDA (Blei et al., 2003b).

• Analyzing topics for queries and integrating topics into data for both the training set and new queries: After the reference set has been analyzed using topic models, it will be used to infer topics for queries. The topic model will then be integrated into the data to build a classifier.

4.1 Query enrichment via click-through links

We automatically extracted click-through links from the query log (which provides us with the title of the image that the user clicks) to enrich the query, represented as a vector $\overrightarrow{q_i}$, with the title of one randomly-chosen click-through associated with it. To further exploit the click-through link, we find the corresponding artwork and extract its keywords: $\overrightarrow{q_i} \cup \overrightarrow{t_i} \cup \overrightarrow{kw_i}$, where $\overrightarrow{t_i}$, $\overrightarrow{kw_i}$ are the vectors of words

[2] http://www.bridgemanart.com

in the title and keywords respectively.

4.2 Hidden Topic Models

The underlying idea is based upon a probabilistic procedure of generating a new set of artworks, where each set refers to titles and keywords of all artworks in a category: First, each set $\overrightarrow{w}_m = (w_{m,n})_{n=1}^{N_m}$ is generated by sampling a distribution over topics $\overrightarrow{\vartheta}_m$ from a Dirichlet distribution ($Dir(\overrightarrow{\alpha})$), where N_m is the number of words in that set m. After that, the topic assignment for each observed word $w_{m,n}$ is performed by sampling a word place holder $z_{m,n}$ from a multinomial distribution ($Mult(\overrightarrow{\vartheta_m})$). Then a word $w_{m,n}$ is picked by sampling from the multinomial distribution ($Mult(\overrightarrow{\varphi}_{z_{m,n}})$). This process is repeated until all K topics have been generated for the whole collection.

Table 1: Generation process for LDA

- M: the total number of artwork sets
- K: the number of (hidden/latent) topics
- V: vocabulary size
- $\overrightarrow{\alpha}$, $\overrightarrow{\beta}$: Dirichlet parameters
- $\overrightarrow{\vartheta}_m$: topic distribution for document m
- $\overrightarrow{\varphi}_k$: word distribution for topic k
- N_m: the length of document m
- $z_{m,n}$: topic index of nth word in document m
- $w_{m,n}$: a particular word for word placeholder [m, n]
- $\Theta = \{\overrightarrow{\vartheta}_m\}_{m=1}^{M}$: a $M \times K$ matrix
- $\Phi = \{\overrightarrow{\varphi}_k\}_{k=1}^{K}$: a $K \times V$ matrix

In order to estimate parameters for LDA (i.e., the set of topics and their word probabilities Φ and the particular topic mixture of each document Θ), different inference techniques can be used, such as variational Bayes (Blei et al., 2003b), or Gibbs sampling (Heinrich, 2004). In this work, we will use Gibbs sampling following the description given in (Heinrich, 2004). Generally, the topic assignment of a particular word t is computed as:

$$p(z_i = k | \overrightarrow{z}_{\neg i}, \overrightarrow{w}) = \frac{n_{k,\neg i}^{(t)} + \beta_t}{[\sum_{v=1}^{V} n_k^{(v)} + \beta_v] - 1} \frac{n_{m,\neg i}^{(k)} + \alpha_k}{[\sum_{j=1}^{K} n_m^{(j)} + \alpha_j] - 1} \quad (1)$$

where $n_{k,\neg i}^{(t)}$ is the number of times the word t is assigned to topic k except the current assignment; $\sum_{v=1}^{V} n_k^{(v)} - 1$ is the total number of words assigned to topic k except the current assignment; $n_{m,\neg i}^{(k)}$ is the number of words in set m assigned to topic k except the current assignment; and $\sum_{j=1}^{K} n_m^{(j)} - 1$ is the total number of words in set m except the current word t. In normal cases, Dirichlet parameters $\overrightarrow{\alpha}$, and $\overrightarrow{\beta}$ are symmetric, that is, all α_k (k = 1..K) are the same, and similarly for β_v (v = 1..V).

4.3 Hidden topic analysis of the Bridgeman metadata

The Bridgeman metadata contains information about artworks in the library that have been annotated by the librarians. We extracted titles and keywords of each artwork, those for which we had a query with a click-through link corresponding to it, and grouped them together by their sub-categories. Each group is considered as a document $\overrightarrow{w}_m = (w_{m,n})_{n=1}^{N_m}$, with the number of total documents M = 732 and the vocabulary size V = 136K words. In this experiment, we fix the number of topics K = 100. We used the GibbsLDA++ implementation[3] to estimate this topic model.

4.4 Building query classifier with hidden topics

Let $Q' = \{\overrightarrow{q_i}'\}_{i=1}^{i=N}$ be the set of all queries enriched via the click-through links, where each enriched query is $\overrightarrow{q_i}' = \overrightarrow{q_i} \cup \overrightarrow{t_i} \cup \overrightarrow{kw_i}$. We also performed Gibbs sampling for all $\overrightarrow{q_i}'$ in order to estimate its topic distribution $\overrightarrow{\vartheta}_i = \{\vartheta_{i,1}, \ldots, \vartheta_{i,K}\}$ where the probability $\vartheta_{i,k}$ of topic k in $\overrightarrow{q_i}'$ is computed as:

$$\vartheta_{i,k} = \frac{n_i^{(k)} + \alpha_k}{\sum_{j=1}^{K} n_i^{(j)} + \alpha_j} \quad (2)$$

where $n_i^{(k)}$ is the number of words in query i assigned to topic k and $n_i^{(j)}$ is the total number of words appearing in the enriched query i.

In order to integrate the topic distribution $\overrightarrow{\vartheta_i} = \{\vartheta_{i,1}, \ldots, \vartheta_{i,K}\}$ into the vector of words $\overrightarrow{q_i}'$ = $\{w_{i,1}, w_{i,2}, \ldots, w_{i,N_i}\}$, following (Phan et al., 2010), we only keep topics whose $\vartheta_{i,k}$ is larger than a threshold *cut-off* and use a *scale* parameter to do the discretization for topics: the number of times topic k integrated to $\overrightarrow{q_i}'$ is round($\vartheta_i \times$ *scale*). After that, we build a Support Vector Machine classifier using SVM light V2.20[4].

[3] http://gibbslda.sourceforge.net/

[4] http://svmlight.joachims.org/

5 Evaluation

In this section, we will describe our training set, gold standard and the performance of our system in comparison with the one in (Le et al., 2011).

5.1 Training set

Manually annotating queries to create a training set in this domain is a difficult task (e.g., it requires the expert to search the query and look at the picture corresponding to the query, etc.). Therefore, we have automatically generated a training set by exploiting a 6-month query log as follow.

First, each query has been mapped to its click-through information to extract the sub-category associated to the corresponding image. Then, from this sub-category, we obtained its corresponding top-cateogry (among the 55 we consider) as defined in BAL taxonomy. The distribution of queries in different categories varies quite a lot among the 55 target categories reflecting the artwork distribution (e.g., there are many more artworks in the library belonging to the category "Religion and Belief" than to the category"Costume and Fashion"). We have preserved such distribution over the target categories when selecting randomly the 15,490 queries to build our training set. After removing all punctuations and stop words, we obtained a training set containing 50,337 words in total. Each word in this set serves as a feature for the SVM classifier.

5.2 Test set

We used the test set of 1,049 queries used in (Le et al., 2011), which is separate from the training set. These queries have been manually annotated by a BAL expert (up to 3 categories per query). Note that these queries have also been selected automatically while preserving the distribution over the target categories observed in the 6-month query log. We call this the "manual" gold standard. In addition, we also made use of another gold standard obtained by mapping the click-through information of these queries with their categories, similar to the way in which we obtain the training set. We call this the "via-CT" gold standard.

5.3 Experimental settings

To evaluate the impact of click-though information and topics in the classifier, we designed the following experiments, where QR is the method without any enrichment and QR-CT-HT is with the enrichment via both click-through and hidden topics.

Setting	Query enrichment
QR	$\overrightarrow{q}$
QR-HT	$\overrightarrow{q} \oplus HT$
QR-CT	$\overrightarrow{q}' = \overrightarrow{q} + \overrightarrow{t} + \overrightarrow{kw}$
QR-CT-HT	$\overrightarrow{q}' \oplus HT$

- $\overrightarrow{q}$: query
- $\overrightarrow{q}'$: query enriched with click-through information
- $\overrightarrow{t}$: click-through image's title
- $\overrightarrow{kw}$: click-through image's keywords
- HT: hidden topics from Bridgeman metadata

Table 2: Experimental Setting

Setting	**Hits**				
	Manual GS				*via-CT*
	# 1	# 2	# 3	$\sum_{Top\_3}$	GS
QR	207	80	24	**311**	**231**
QR-HT	212	81	25	**318**	**235**
QR - CT	243	107	38	**388**	**266**
QR - CT - HT	289	136	49	**474**	**323**

Table 3: Results of query classification: number of correct categories found (for 1,049 queries)

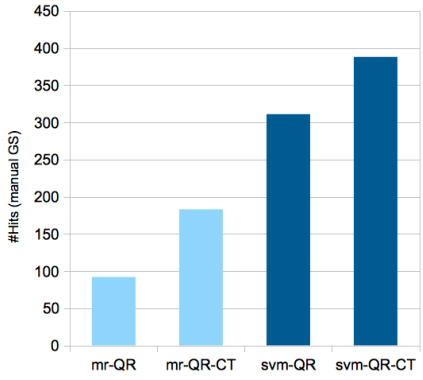

Figure 1: The impact of click-through information with matching-ranking (mr) and our approach (svm)

To answer our first research question, namely whether click-through information and hidden topics are useful for this query classifier, we examine the number of correct categories found by the classifier built both with and without the enrichment. The results of the experiment are reported in Table 3. As can be seen from the table, we notice that the click-through information plays an important role. In par-

ticular, it increases the number of correct categories found from 311 to 388 (compared with the *manual* GS) and from 231 to 266 (using the *via-CT* GS).

To answer our second research question, namely whether integrating the enriched information into a machine learning classifier can perform better than the matching and ranking method, we also compare the results of our approach with the one in (Le et al., 2011). Figure 1 shows the impact of the click-through information for the SVM classifier (svm) in comparison with the matching and ranking approach (mr). Figure 2 shows the impact of the hidden topics in both cases. We can see that in both cases our classifier outperforms the matching-ranking one considerably (e.g., from 183 to 388 correct categories found in the QR-CT-HT method).

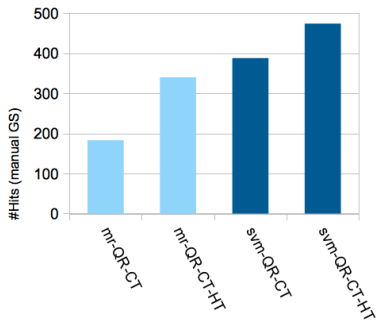

Figure 2: The impact of hidden topics with matching-ranking (mr) and our approach (svm)

However, in the case where we use only queries without click-through information, we can see that hidden topics do not bring a very strong impact (the number of correct categories found only slightly increases by 7 - using the "manual" gold standard). The result might come from the fact that this topic model was built from the metadata, using only click-through information, but has not been learned with queries.

6 Conclusion

In this study, we have presented a machine learning classifier for query classification in an art image archive. Since queries are usually very short, thus difficult to classify, we first extend them with their click-through information. Then, these queries are further enriched with topics learned from the BAL metadata following (Le et al., 2011). The result from this study has confirmed again the effect of click-through information and hidden topics in the query classification task using SVM. We have also described our method of automatically creating a training set based on the selection of queries mapped to the click-through links and their corresponding available categories using a 6-month query log. The result of this study has shown a considerable increase in the performance of this approach over the matching-ranking system reported in (Le et al., 2011).

7 Future work

For future work, we are in the process of enhancing our experimentation in several directions:

Considering more than one click-through image per query: In this work, we have considered only one category per query to create the training set, while it might be more reasonable to take into account all click-through images of a given query. In future work, we plan to enrich the queries with either all click-through images or with the most relevant one instead of randomly picking one click-through image. In many cases, a click-through link is not necessarily related to the meaning of a query (e.g., when users just randomly click on an image that they find interesting). Thus, it might be useful to filter out those click-through images that are not relevant.

Enriching queries with top hits returned by the BAL search engine: In the query logs, there are many queries that do not have an associated click-through link. Hence, we plan to exploit other enrichment method that do not rely on those links, in particular we will try to exploit the information coming from the top returned hits given by the library search engine.

Analyzing queries in the same session: It has been shown in some studies (Cao et al., 2009) that analyzing queries in the same session can help determine their categories. Our next step is to enrich a new query with the information coming from the other previous queries in the same session.

Optimizing LDA hyperparameters and topic number selection: Currently, we fixed the number of topics $K = 100$, the Dirichlet hyperparameters $\alpha = 50/K = 0.5$ and $\beta = 0.1$ as in (Griffiths and

Steyvers, 2004). In the future, we will explore ways to optimize these input values to see the effect of different topic models in our query classification task.

Exploiting visual features from the BAL images: The BAL dataset provides an interesting case study in which we plan to further analyze images to enrich queries with their visual features. Combining text and visual features has drawn a lot of attention in the IR research community. We believe that exploiting visual features from this art archive could lead to interesting results in this specific domain. A possible approach would be extracting visual features from the click-through images and representing them together with textual features in a joint topic distribution (e.g., (Blei et al., 2003a; Li et al., 2010)).

Comparing system with other approaches: In the future, we plan to compare our system with other query classification systems and similar techniques for query expansion in general. Furthermore, the evaluation phase has not been carried out thoroughly since it was difficult to compare the one-class output with the gold-standard, where the number of correct categories per query is not fixed. In the future, we plan to exploit the output of our multi-class classifier to assign up to three categories for each query and compute the precision at n.

Acknowledgments

This work has been partially supported by the GALATEAS project (http://www.galateas.eu/ – CIP-ICT PSP-2009-3-25430) funded by the European Union under the ICT PSP program.

References

Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, and David Grossman. 2005. Automatic web query classification using labeled and unlabeled training data. In *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 581–582. ACM Press.

David M. Blei, Michael I, David M. Blei, and Michael I. 2003a. Modeling annotated data. In *In Proc. of the 26th Intl. ACM SIGIR Conference*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003b. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Andrei Z. Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. 2007. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 231–238, New York, NY, USA. ACM.

Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum*, 36:3–10, September.

Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxi Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. 2009. Context-aware query classification. In *SIGIR'09, The 32nd Annual ACM SIGIR Conference.*

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1(Suppl 1):5228–5235.

Gregor Heinrich. 2004. Parameter estimation for text analysis. Technical report.

Dieu-Thu Le, Raffaella Bernardi, and Edwin Vald. 2011. Query classification via topic models for an art image archive. In *Recent Advances in Natural Language Processing, RANLP, Bulgaria*.

Li-Jia Li, Chong Wang, Yongwhan Lim, David Blei, and Li Fei-Fei. 2010. Building and using a semantivisual image hierarchy. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, June.

Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha. 2010. A hidden topic-based framework towards building applications with short web documents. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints).

Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Giang Yang. 2006a. Query enrichment for web-query classification. *ACM Transactions on Information Systems*, 24(3):320–352.

Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2006b. Building bridges for web query classification. In *SIGIR'06*.