Wikipedia based Unsupervised Query Classification

Milen Kouylekov, Luca Dini, Alessio Bosca, and Marco Trevisan

CELI S.R.L., Torino, Italy
{kouylekov,dini,bosca,trevisan}@celi.it

Abstract. In this paper we present an unsupervised approach to Query Classification. The approach exploits the Wikipedia encyclopedia as a corpus and the statistical distribution of terms, from both the category labels and the query, in order to select an appropriate category. We have created a classifier that works with 55 categories extracted from the search section of the Bridgeman Art Library website. We have also evaluated our approach using the labeled data of the KDD-Cup 2005 Knowledge Discovery and Data Mining competition (800,000 real user queries into 67 target categories) and obtained promising results.

Keywords: Query Classification, Wikipedia, Vector Models

1 Introduction

In Information Science logs analysis and more specifically Query Classification (QC) has been used to help detecting users web search intent. Query classification studies have shown the difficulty of achieving accurate classification due to the inevitably short queries. A common practice is to enrich query with external information, and to use an intermediate taxonomy to bridge the enriched query and the target categories. A good summary of such approaches is made by the KDDCUP 2005 organizers [2005]). Associating external information to queries is costly as it involves crowing the web. The goal of our approach is to create an lightweight unsupervised language independent approach to query classification using the rich content provided by the Wikipedia, a resource easily accessible in many languages.

In Section 2 we present this approach. In Section 3 we provide a evaluation of its capabilities.

2 Unsupervised Query Classification Approach

Our approach is based on the vector space model (VSM). The core of the VSM is representing text documents as vectors of identifiers, such as index terms. Each element of the vectors corresponds to a separate term found in the document. If a term occurs in the document, its value in the vector is non-zero. The dimensionality of the vector is the number of words in the vocabulary (the number

of distinct words occurring in the corpus). In VSM, weights associated with the terms are calculated based on a heuristic functions. Some of the more popular approaches are term frequency and inverse document frequency.

Using the VSM a similarity function between the vectors of two documents and a query can be defined. The standard function used is the cosine similarity coefficient, which measures the angle between two vectors.

We have adapted the vector space model to the query classification task using the following approach: First, we associate a document that describes the category with each category C_k. We name this document category document CD. For example a CD for the category Basketball must contain information about: the rules of the game, *National Basketball Association*, *FIBA* and famous players etc. A CD for the Arts must contain information about i) paining; ii) sculptures; 3) art museums etc. In the second stage of the approach we associate to each query a set of relevant documents found in a document collection. We define the category score of query q for category C_k as the maximum value of the cosine similarity between term vectors of the CD of the category and a document relevant to the query ${}_Qi$. For example we expect to find a lot of common terms between a document relevant to the query *Michael Jordan* and the CD for category Basketball and few common terms with the CD of the category Art. Finally as output the approach returns the categories which CD has the highest cosine similarity with a relevant document of the query.

We use relevant documents for a query and not the query because we compare the query terms with CD documents that are not relevant to the query itself. For example the CD of the category Arts does not contain a mentioning of Pablo Picasso but the intersection between it and documents relevant to Pablo Picasso contain a lot of common terms like painting, art, surrealism etc.

In order to make our approach feasible we need a document collection that contains sufficient number of documents in order to: 1) Find a big enough CD document for each category. 2) Find documents relevant to the classified queries.

The advantage of the proposed approach is that it does not require training data and it is language dependent. The approach will befit greatly from a short category description as this will allow a more correct selection of a CD.

3 Experiments

In our experiments we used the Wikipedia, a free, web-based, collaborative, multilingual encyclopedia[1]. We assign as CD for the category with the name X the Wikipedia page with the same title. For example the Wikipedia Page with title *Basketball* can be used as a CD for the category *Basketball*. The page describes almost all the important aspect of the game and has a lot of terms in common with documents relevant to queries like: *Michael Jordan*, *NBA Playoffs* and *Chicago Bulls*. Respectively the page with the title *Hardware* contains a lot of terms in common with documents relevant for queries like: *intel processors*, *computer screens* and *nvida vs intel*.

[1] http://wwww.wikipedia.org

For some categories the CD assigned by the approach contains short texts that are not sufficient for a complete overview of the category. We expand the CD for these categories by concatenating the texts of the pages that contain the name of the category as sub part of the page title. For example the CD for the category *Arts* can be expanded by concatenating the text of the pages: *Liberal arts*, *Visual arts*, *Arts College*, *Art Education*, *Islamic Arts* etc. If the approach does not find a page with the same title as the category it assigns as CD for the category the concatenation of pages with the name of the category as sub part of the page title or pages that contain the category name in the first sentence of the page text.

3.1 Bridgeman Art Library

Our first evaluation is done using the taxonomy and query dataset created for the ART domain in the Galateas Project [2011]. The domain is defined by the contents of the Bridgeman Art Library(BAL) website[2]. To understand their use and meaning the categories have been grouped by BAL domain experts into three groups: Topics (Land and Sea, Places, Religion and Belief, Ancient and World Cultures etc. 23), Materials (Metalwork, Silver, Gold & Silver Gilt, Lacquer & Japanning, Enamels etc. 10), and Objects (Crafts and Design, Manuscripts, Maps, Ephemera, Posters, Magazines, Choir Books etc. 22) (total: 55 top-categories).

Our approach was evaluated on the 100 queries annotated by the three annotators into upto 3 categories. The queries were in four languages English , French, German, Dutch and Italian. We have created a classification instance for each language by manually translating the name of the categories in each language. For each of these queries we automatically assign the top 3 categories returned by the classifier. Example:

Query: navajo turquoise
Category1: Semi-precious Stones Score: 0.228
Category2: Silver, Gold & Silver Gilt: 0.1554
Category3: Botanical Score: 0.1554

In this example the category assigned to the query is Category1 *Semi-precious Stones*. The results of the evaluation are summarized in Table 1.

	Precison	F-Measure
Bridgeman Art Library	16.1	14.5
KDD Cup Results	29.0	32.2

Table 1. Results

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

3.2 KDD Cup 2005

To evaluate our approach we have experimented also with the KDD-Cup 2005 data [2005]. The data is from query classification task selected by the organizers as interesting to participants from both academia and industry. The task of the competition consists in classifying Internet user search queries. The participants had to categorize 800,000 queries into 67 predefined categories. The meaning and intention of search queries is subjective. A search query *Saturn* might mean *Saturn car* to some people and *Saturn the planet* to others. The participants had to tag each query with up to 5 categories. The systems participating in the competition were ranked by the organizers on the obtained average **F-Measure**.

We have evaluated our approach against a gold standard provided by the organizers (800 queries). Our approach obtained an average F-Measure of 32.2 (Table 1). The state of the art system [2006] in the competition achieves F-Measure of 46.1.

4 Discussion

The results obtained are encouraging having in prospective the unsupervised nature of the approach. One of the main difficulties were the generic names of categories like *Icons* and *The Arts and Entertainment*. The documents for these categories contained terms that were not relevant to the art domain. Also a significant problem for the system posed queries that are named entities. These queries were classified based on their descriptions into categories relevant to their peculiarities and not in *People and Society*, *Personalities* and *Places* categories. A possible solution to this problem will be to map DBPedia[3] hierarchy to the domain categories and use it as a additional source of knowledge.

The results we obtained is encouraging particularly because we did not associated additional information to each queries apart of the documents obtained using Wikipedia. Many of the queries did not produce relevant Wikipedia documents, which is one of the main limitations of our approach. Additional domain corpora will decrease its effect.

References

[2005] Ying Li, Zijian Zheng, and Honghua (Kathy) Dai. Kdd cup-2005 report: facing a great challenge. ACM SIGKDD Explorations Newsletter Homepage archive, 7(2), 2005.

[2006] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. In SIGIR06, 2006.

[2011] Eduard Barbu, Raphaella Bernardi, T.D. Le, Milen Kouylekov, V. Petras, Massimo Poesio, Juliane. Stiller, E. Vald, D7.1 First Evaluation Report of Topic Computation and TLIKE , Galateas Project http://www.galateas.eu

[3] http://dbpedia.org/