Information Theoretic Approach to Information Extraction

Giambattista Amati

Fondazione Ugo Bordoni
Rome Italy

Abstract. We use the hypergeometric distribution to extract relevant information from documents. The hypergeometric distribution gives the probability estimate of observing a given term frequency with respect to a prior. The lower the probability the higher the amount of information is carried by the term. Given a subset of documents, the information items are weighted by using the inversely related function of of the hypergeometric distribution. We here provide an exemplifying introduction to a topic-driven information extraction from a document collection based on the hypergeometric distribution.

1 Introduction

Information extraction is a wide concept denoting the art of extracting tokens, items or any other form of *relevant* textual information from a document collection. The basis of information extraction is the notion of relevant item with respect to the information need or task. Extracted information can be conceived either as a self-contained, or distilled text that carrying most of the information content that is however sufficient to explain, answer or complete the information need or task in a minimal, or compact or structured form. For example data extracted from documents may be regarded as attribute values for record fields to be filled into a relational database, like for example the undirected graphical models of the conditional random fields [11, 18]. A primary goal of information extraction is to locate a minimal set containing most of the candidate items for extracting information from the data. Once that the relevant information is extracted, it is further possible to classify and insert candidate items into predefined relations, categories or other structured textual information. This paper deals with the problem of circumscribing relevant information conditioned to an information need or task employing a very low cost of computational time and space. We indeed introduce a topic-biased selection of relevant items for information extraction. This selection problem is similar to the construction of automatic query expansion, and it can be indeed adapted to many different information retrieval and extraction tasks.

The items extracted with our technique torn out of their topic contexts might be without a self-contained meaning. We treat the extracted items as atomic elements or basic constituents to be used to single out other relevant pieces of text, whether they are attribute values, or phrases, paragraphs and sections. These pieces of text have the highest *information content* with respect to the background knowledge (for example a topic $\mathbf{q}$), and if they are not used directly for filling field records of a relational database, then they can be used to weight and extract larger pieces of information.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 519–529, 2006.

To perform information extraction, we first define what in probability theory is called the population of the background knowledge $\mathbf{q}$. To define the topic population, we gather relevant documents into a single document set $\mathcal{P}_{\mathbf{q}}$. *This document set is automatically obtained as the set of pseudo-relevant documents, that is the set of the topmost retrieved documents given the query* $\mathbf{q}$. Then we use the pooling set $\mathcal{P}_{\mathbf{q}}$ to sample frequencies of terms. Thus, for each term of the population we compute the probability $\text{Prob}(\mathbf{f}|\mathbf{p}, \mathcal{P}_{\mathbf{q}})$ of the observed term frequency $\mathbf{f}$ in the pool $\mathcal{P}_{\mathbf{q}}$ with respect to the prior probability $\mathbf{p}$ of the term, where Prob is the hypergeometric distribution [1]. The *information content* of the item $\mathbf{t}$ conditioned to the background knowledge $\mathbf{q}$ is defined as the inverse relation of the probability:

$$\text{Inf}(\mathbf{t}|\mathbf{q}) = -\log \text{Prob}(\mathbf{f}|\mathbf{p}, \mathcal{P}_{\mathbf{q}})$$

We finally weight and select all terms $\mathbf{t}$ with the highest $\text{Inf}(\mathbf{t}|\mathbf{q})$. This selection procedure is actually very efficient, because $\text{Inf}(\mathbf{t}|\mathbf{q})$ only requires few accesses to small portions of the direct file of the collection. The compressed direct file of the collection occupies the same space of the compressed inverted file and decompression time of the loaded parts is irrelevant, because file access and processing concern only few documents per query.

Note that Prob is a second-order probability: it is the probability of obtaining a probability $\mathbf{f}$ given a prior $\mathbf{p}$. Therefore, we do not maximise the likelihood $\text{Prob}(\mathbf{f}|\mathbf{p}, \mathcal{P}_{\mathbf{q}})$ to extract a probability value for $\mathbf{p}$ as in the maximum likelihood estimator (MLE), but we directly use the probability Prob as an estimator of information content of the term $\mathbf{t}$.

To exemplify our methodology we now use three topics as examples. The collection is from disks 4 and 5 of TREC minus the CR collection and consists of about 2 Gbytes of data, with 528,107 documents. The three examples are non-informational topics of TREC. Instead these three topics require a list of values or names to be associated to the query.

We use two parameters to measure the specificity of the returned answer: the number of documents of the pool $\mathcal{P}_{\mathbf{q}}$, which is 8, and the minimal number of documents of the pool containing the term. If a term belongs at least to one document of $\mathcal{P}_{\mathbf{q}}$, then we call the term to be of *level 1 of generality*. It is of *level 2* if it belongs at least to 2 documents, and so on. As soon as the level n of generality increases the high informative terms become more general, being frequent but in several different documents of the pool. The level of generality of a term in the pool is a very important indicator of the semantic role of the term in the documents. If a term is in the level 1 but not in the level 2, then it is high probable that the term is an attribute value of some record field. On the contrary, as soon as the level of generality of the term is high, then the informative term might be used as a relation or a concept in a relation database. The examples below will illustrate this issue.

Example 1. Nobel prize winners. Among the first 37 new retrieved terms at level 1 we have 8 Nobel prize winners (they are in italics) out of 11 proper names (words that are not in dictionary[1]): prize (0.4953), nobel (0.1577), winner (0.1245),*MacRobert*

[1] We also provide an automatic method able to detect correct proper names without using dictionary or other natural language analysis in Section 4.