Evaluation Methods for Focused Crawling

Andrea Passerini, Paolo Frasconi, and Giovanni Soda

DSI, University of Florence, ITALY
{passerini,paolo,giovanni}@dsi.ing.unifi.it

Abstract. The exponential growth of documents available in the World Wide Web makes it increasingly difficult to discover relevant information on a specific topic. In this context, growing interest is emerging in *focused crawling*, a technique that dynamically browses the Internet by choosing directions that maximize the probability of discovering relevant pages, given a specific topic. Predicting the relevance of a document before seeing its contents (i.e., relying on the parent pages only) is one of the central problem in focused crawling because it can save significant bandwidth resources. In this paper, we study three different evaluation functions for predicting the relevance of a hyperlink with respect to the target topic. We show that classification based on the anchor text is more accurate than classification based on the whole page. Moreover, we introduce a method that combines both the anchor and the whole parent document, using a Bayesian representation of the Web graph structure. The latter method obtains further accuracy improvements.

1 Introduction

The World Wide Web is experiencing an exponential growth, both in size and number of users. The quantity and variety of documentation available poses the problem of discovering information relevant to a specific topic of interest. The instruments developed to easen information recovering in the Internet suffer from various limitations. *Web directories* cannot realize exhaustive taxonomies and have a high mantainance cost due to the need for human classification of new documents. *Search engines* allow only searches by keywords, and cannot compete with dynamysm of the Internet in terms of coverage, novelty, and consistence of information [6]. These limitations suggest to experience different solutions, trying to provide focused, consistent and possibly new information related to a specific topic of interest. *Focused crawling* is a techinque that dynamically browses the Web looking for documents relevant to a certain topic. It employs an evaluation method to choose the best hyperlink to follow at a given time in order to maximize the probability of discovering relevant information, given a topic of interest. This allows to fully exploit limited bandwidth resources. The evaluation methods deal with the problem of predicting relevance of a document without seeing its content, while knowing the contents of a page pointing to the document to be evaluated. A common way to implement this evaluation function is that of training a classifier to predict the relevance of a document to a specific topic, and using such a relevance, calculated for the visited page, as a score for all the hyperlinks contained in the page [2, 3]. This method, which we shall call *neighbourhood* score, relies on the idea that pages treating a specific topic will point to other pages relevant to that topic. More complex evaluation functions

assign each link a different score that depends on the information contained in the context surrounding the hyperlink inside the page [5, 7].

Focused crawling is a rather young technique, and almost no comparisons are available for the different evaluation methods employed. In this paper, we propose a method to compare predicting capabilities of different evaluation functions. We use such a method to compare the neighobourhood score with an evaluation function which assigns each link a different score that depends on their anchor text. The *anchor* score greatly outperforms the neighbourhood score in every experiment performed. We also introduce a method that combines the other two ones, exploiting a Bayesian representation of hypertext connections between documents. Such a method yields further improvements in terms of predicting capability.

2 The Probabilistic Model for Hypertexts

A typical representation of textual documents in information retrieval is the so called *bag of words*. A document is represented as in a vector space whose dimension is the vocabulary size. Vector components are proportional to word frequencies in the document being represented. We restrict to alphabetical words and remove stopwords. Stemming and feature selection did not prove to be effective in this task. Hypertext documents have a structured nature given by HTML tags. We try to mantain part of this structure by using an extension of the bag of words representation. We split the document in three areas associated with HTML tags: 1) META, 2) TITLE and H1, 3) all the remaining tags. A different vocabulary was used for each area.

Naive Bayes is a well known approach to text categorization. The basic underlying assumption is that words in the document are conditionally independent given the class. The associated Bayesian network is shown in Figure 1a, where X_i is the i-th word in the document, and C is the document class. Maximum a posteriori predictions are obtained as

$$c^* = \arg\max_{c_i \in \mathcal{C}} P(c_i) \prod_{j=0}^{|d|} P(w_j|c_i) \tag{1}$$

where $\mathcal{C}$ is the set of candidate classes, w_j are the words in document d ,and $|d|$ is the document length. Given a set of training examples for each class, the a priori class probabilities $P(c_i)$ are obtained as the number of training documents belonging to each class divided by the total number of training documents, while the probabilities associated to words are given by:

$$P(w_j|c_i) = \frac{n_{ji} + 1}{n_i + |V|} \tag{2}$$

where n_{ji} is the number of times word w_j occurs in documents of class c_i, n_i is the total number of words occurrences in documents of class c_i, and $|V|$ is the size of the vocabulary.

We suggest an extension of the Naive Bayes classifier in order to exploit the structured representation given by HTML tags. We associate multiple random variables with a word, depending on the tag context (TC) in which it can be

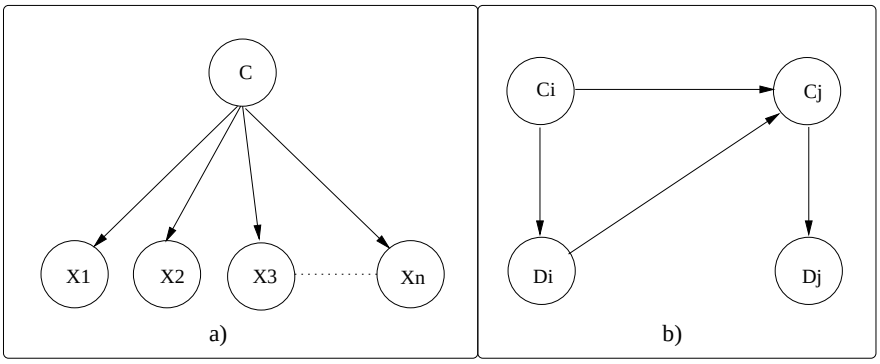

Fig. 1. a) Structure of a Naive Bayes classifier. b) Bayesian model for two linked documents.

found inside a document, and use a different vocabulary for each tag context. The class probability is calculated as:

$$c^* = \arg\max_{c_i \in C} P(c_i) \prod_{k=1}^{m} \left(\prod_{j=0}^{|d_{V_k}|} P(w_{d_{V_k} j} | c_i) \right) \tag{3}$$

where m is the number of TC being considered, d_{V_k} the fraction of words in d belonging to TC V_k, and $w_{d_{V_k} j}$ the j^{th} of these words. Probabilities associated with words are given by:

$$P(w_{kj} | c_i) = \frac{n_{kji} + 1}{n_{ki} + |V_k|} \tag{4}$$

where n_{kji} is the number of occurrences of word w_j in the k^{th} TC of documents of class c_i, n_{ki} is the total number of words occurrences in the k^{th} TC of documents of class c_i, and $|V_k|$ is the size of the k^{th} vocabulary.

This classifier proved to perform better than regular Naive Bayes in classifying HTML documents, and we shall employ it in developing the evaluation functions for hyperlinks.

3 Evaluation Methods for Hyperlinks

In order to assign a relevance score to hyperlinks inside a document, we must implement an evaluation function that predicts the probability that the document pointed by an hyperlink belongs to the class of interest. While different evaluation functions have been proposed, no significant comparison experiments are available in the literature. In this paper, we compare three different evaluation fuctions, that we named: *Neighbourhood* score, *Anchor* score, *Transition* score.

The first method we implement is the simplest and commonest one. It assigns to all hyperlinks inside a document the same relevance, given by the class a posteriori probability of the document containing them. We call this method *neighbourhood* score, because it is based on the assumption that relevant pages for a given class will point to other relevant pages for the same class. We calculate such a probability by the extended Naive Bayes classifier described in previous section.

The second method, called *anchor* score, assigns each hyperlink a different score depending on the text of its anchor, that is the text that can be clicked when viewing the document with a browser, or the text contained in the ALT tag in case of a clickable image. We employ a Naive Bayes classifier trained on the anchor text of hyperlinks whose pointed page class was known. This method is opposite to the neighbourhood one because it doesn't take in account neither the remaining text of the document containing the hyperlink to classify nor its class.

The third method, called *transition* score, aims to merge the two contributions of neighbourhood and anchor score into a single evaluation function. In this case, we use a Bayesian network model for the relationship between two documents connected by a hyperlink. The model is shown in Figure 1b, where D_i is a document, with associated class C_i, which contains a hyperlink pointing to a document D_j with associated class C_j. We are interested in estimating information related to edge $D_i \rightarrow C_j$, that is probability that a hyperlink in D_i points to a document of class C_j, when C_j is the class of interest. Named D_{ij} a hyperlink inside D_i, we can represent the probability that the class of the document pointed by that link is C_j as:

$$P(C_j|D_{i,j}) = \sum_{C_k \in C} \left(\frac{P(D_{i,j}|C_j, C_k)P(C_j|C_k)P(C_k|D_{i,j})}{P(D_{i,j}|C_k)} \right) \tag{5}$$

where probability is summed over all the possible classes $C_k \in C$ of D_i, and we applied product rule and then Bayes theorem to the first term of the product. Unfortunately equation 5 has too many parameters to be estimated, and some simplifying assumption is necessary. We use the assumption that the link is independent of the class of the document containing it, that is $D_{i,j} \perp C_k$. With such an assumption the equation becomes:

$$P(C_j|D_{i,j}) = \frac{P(D_{i,j}|C_j) \sum_{C_k \in C} P(C_j|C_k)P(C_k|D_{i,j})}{P(D_{i,j})} \tag{6}$$

In the above equation, $P(C_k|D_{i,j})$ is the probability of class C_k given link $D_{i,j}$. This is the neighbourhood score contribution, i.e. the a posteriori probability of the class given the document in which the link is contained. $P(C_j|C_k)$ represents the *transition* probability, that is the probability that a generic document of class C_k has a link pointing to a document of class C_j. Finally, $P(D_{i,j}|C_k)$ is the probability that a link $D_{i,j}$ is contained in a document of class C_k. $P(D_{i,j})$ is the a priori probability of a link $D_{i,j}$.

4 Experimental Results

In order to compare performances of the proposed evaluation methods, we developed a dataset of documents containing labeled hyperlinks, i.e. hyperlinks pointing to pages whose class was known. To generate such a set, we started from a dataset provided by the *World Wide Web Knowledge Base*[1] project, consisting

[1] the dataset is available on-line at:
http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data

of pages from computer science departments of United States Universities. We merged together the classes referring to personal homepages (faculty members, staff and students) obtaining five classes: *Course* for university courses, *Department* for department homepages, *Homepage* for personal homepages, *Project* for research projects and *Other* for pages not belonging to any of the other classes. We then collected all the *backlinks* of these pages, that is documents containing links pointing to one of such pages. In order to have more than one labeled hyperlink per document, we also collected part of the other links contained inside backlink documents, and manually classified the corresponding pages. In this way, each example consists of a labeled hyperlink and the document containing the link itself.

Class	0	1	2	3	4	Class Total	Recall (%)
0 course	217	.	851	208	4	1280	16.95
1 dept.	56	2	1233	197	54	1542	0.13
2 home	102	1	1818	754	748	3423	53.11
3 other	96	3	2542	1039	52	3732	27.84
4 project	8	.	508	281	55	852	6.46

Neighbourhood score (**28.91**% accuracy)

Class	0	1	2	3	4	Class Total	Recall (%)
0 course	882	101	37	78	182	1280	68.91
1 dept.	117	1303	32	26	64	1542	84.50
2 home	261	161	2560	334	107	3423	74.79
3 other	476	266	139	2040	811	3732	54.66
4 project	63	42	17	86	644	852	75.59

Anchor score (**68.60**% accuracy)

Class	0	1	2	3	4	Class Total	Recall (%)
0 course	673	92	73	255	187	1280	52.58
1 dept.	39	1302	55	99	47	1542	84.44
2 home	43	114	2779	368	119	3423	81.19
3 other	165	283	180	2402	702	3732	64.36
4 project	17	33	32	161	609	852	71.48

Transition score (**71.71**% accuracy)

Table 1. Total Accuracies and Confusion matrices (row is actual, column is predicted).

We divided the examples in a training set of 35,606 examples and a test set of 10,829 examples. For each evaluation method proposed, we trained the corresponding classifier on the training set and verified its performances on the test set. The following tables show the results in accuracy and recall, together with the confusion matrices, for the three methods proposed (table 1). Anchor score outperforms neighbourhood score of 39.69%, meaning that, when present, anchor text alone gives much better information about the page pointed by the link than the document cointaining the link itself.

Adding transition probabilities yields a 10% prediction error reduction. Modeling the probability of transition between classes, weighted by the a posteriori probability of the class of the starting document, helps to disambiguate in the case of rare hyperlinks, like for example a department homepage directly pointing to a specific course homepage.

5 Conclusions

In this paper, we showed that an evaluation function based on anchor text can greatly outperform the common approach of assigning the same score to all links contained in a given page.

We furthermore proposed an evaluation function exploiting a Bayesian representation of connection between documents, showing that it further increases predicting accuracy. The problem of this approach is that anchor score does not contribute to the evaluation when the anchor text is empty. A further development is to extend the text used for anchor score to some context of the hyperlink itself. Using context for hyperlink score has been proposed by previous works [7], but there is no evidence of its benefit, especially for the difficulty to define and extract a consistent context without introducing greater noise. We are trying to apply the idea of using a link context path [1] to extract such a context. Regarding focused crawling, we are studying techniques to choose directions to follow when no relevant documents are available in the fringe of search, and longer distance predictions must be made [4, 7].

References

1. G. Attardi, S. Di Marco, and D. Salvi. Categorization by context. *Journal of Universal Computer Science*, 4(9):719–736, 1998.
2. P. De Bra, G.-J. Houben, Y. Kornatzky, and R. Post. Information retrieval in distributed hypertexts. In *Proceedings of RIAO'94, Intelligent Multimedia, Information Retrieval Systems and Management*, New York, NY, 1994.
3. S. Chakrabarti, M. van der Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proceedings of the 8th International World Wide Web Conference*, Toronto, Canada, 1999.
4. M. Diligenti, F.M. Coetzee, S. Lawrence, C.L. Giles, and M. Gori. Focused crawling using context graphs. In *Proceedings of the 6th International Conference on Very Large Databases, VLDB 2000*, Cairo, Egypt, 2000.
5. M. Hersovici, M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalheim, and S. Ur. The shark-search algorithm — an application: tailored web site mapping. In *Proceedings of the 7th International World Wide Web Conference (WWW7)*, Brisbane, Australia, 1998.
6. S. Lawrence and C.L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, July 1999.
7. J. Rennie and A. McCallum. Using reinforcement learning to spider the web efficiently. In *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, 1999.