Web page categorization without the web page

Min-Yen Kan
Department of Computer Science, National University of Singapore
3 Science Drive 2, Singapore 117543
kanmy@comp.nus.edu.sg

ABSTRACT

Uniform resource locators (URLs), which mark the address of a resource on the World Wide Web, are often human-readable and can hint at the category of the resource. This paper explores the use of URLs for web page categorization via a two-phase pipeline of word segmentation/expansion and classification. We quantify its performance against document-based methods, which require the retrieval of the source document.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**] Content Analysis and Indexing – *Linguistic processing*

Keywords: Uniform Resource Locator, word segmentation, abbreviation expansion, text categorization

1. INTRODUCTION

Web page indices only cover a fraction of the accessible web. All of these resources are specified in terms of their *Universal Resource Locator* (URL), a string that specifies a protocol, a host and path to locate the resource.

Web indexers typically work by first retrieving a document and processing it. This process often yields the URLs of additional documents that can be retrieved and processed in subsequent iterations. As web resources are often hyperlinked to more than a single page, this process creates a growing list of documents to be spidered – documents whose URLs are known but which have not yet been retrieved and processed. Work on focused crawling can partially help in addressing this bottleneck (e.g., [1]) by addressing areas that need more coverage. Techniques that examine just the URLs can also target this bottleneck, as the source document need not be retrieved.

Often the URL itself is quite informative, as human experts can glean a large amount of information from it without needing to examine the actual contents of the resource. An example URL shows this intuition:

http://cs.cornell.edu/Info/Courses/Current/CS415/CS414.html

Given categories such as course, faculty, project, and student, it is easy to guess that the page belongs to the course category. Whereas much published work on categorization ignores the URL as a source of information, we examine how to make maximal use of this single resource. Our system thus performs web page classification using only the URL in a two-stage process.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

2. SEGMENTING URLS

A URL is first divided to yield a *baseline* segmentation: its components as given by the URI protocol (e.g., *scheme :// host / path-elements / document . extension*), and further segmented wherever one or more non-alphanumeric characters appear (e.g., *faculty-info* → *faculty info*). We further break the baseline segments with some processing to arrive at a *refined* segmentation: if a transition between uppercase, lowercase and digits was observed. Two approaches were used to break the refined segments down further.

Information content reduction uses information content as a criterion for splitting. This algorithm examines all $2^{||c||}$ possible partitions of the chunk c and calculates the sum of the information content (IC) of each partitioning, defined as its negative log probability of occurrence, $-log(p(x))$. To calculate the IC for partition elements, their probability is needed. We estimate such probabilities using data from the WebBase project, which provides the document frequency of tokens over 39 million web pages.

A partitioning that has a lower IC sum than other partitionings can be said to have a lower amount of uncertainty, and is a more probable parse of the chunk. A similar approach has been applied [2] to Chinese word segmentation. The system finds the partition with the minimal IC and compares it with the IC of the string as a series of characters (using unigram character probabilities). If the minimum scoring partition has lower information content than the chunk as a series of characters, then the chunk is further broken down into the partition's segments. Otherwise the chunk is kept as a single segment.

Title token based finite state transducer (FST) tries to simultaneously split and expand segments based on previously-seen web page titles. Consider the URL fragment "cs", which might correspond to "computer science" in a majority of the training pages' title in which it appears. If the same URL fragment "cs" is encountered in the testing corpus, it is automatically expanded to "computer science".

To associate a fragment with a sequence of title words in the training corpus, a weighted non-deterministic finite-state transducer is employed. The transducer has a small set of rules that associate a score with certain moves that match or skip letters in the title tokens with corresponding letters in the segment. An expansion in the must cover all letters in the segment to be considered valid. The expansion or segmentation that scores highest is used as the expansion for a particular instance. The rules that are used are listed in Table 1.

As an example, given the computer *nytimes* and title tokens "New York Times", the FST selects the following series of transitions: (ø $\xrightarrow{R_1}$ N $\xrightarrow{R_5}$ e $\xrightarrow{R_5}$ w $\xrightarrow{R_1}$ Y $\xrightarrow{R_5}$ o $\xrightarrow{R_5}$ r $\xrightarrow{R_5}$ k $\xrightarrow{R_1}$ T $\xrightarrow{R_3}$ i $\xrightarrow{R_3}$ m $\xrightarrow{R_3}$ e $\xrightarrow{R_4}$ s), as it has the maximal score 12, and outputs $|n|y|times$. Then, if the

FST Rule	Score	Output
1. Match the initial letter in the subsequent token	2	$\|l$
2. Match the initial letter in a non-subsequent token	1	$\|l$
3. Match a subsequent letter in the current token	1	l
4. Match the final letter in the current token	3	l
5. Skip a character in the candidate expansion	0	ϵ

Table 1: Rules for the title expansion weighted FST.

URL fragment "nytimes" is encountered in the testing corpus, the words "new york times" would be substituted for the fragment.

3. CATEGORIZING EXPANSIONS

The resulting set of tokens is fed to the SVM^{light} machine learner to classify. As SVM^{light} needs numeric features and not lexical tokens, each unique token is assigned an ID and the simple counts of the token in the set are used as the feature's value.

To test the effectiveness of the system, we evaluate on the WebKB corpus, commonly used for web classification experiments. The WebKB corpus consists of web pages collected from four universities, classified into seven categories. We employ a subset of the WebKB, containing 4,167 pages (the ILP 98 dataset [3]), in which each page is associated with its anchor words (text from the hyperlinks that point to the page). The classification was set up as a series of binary classification tasks, done in the same manner as [4]: using only the *student*, *faculty*, *course* and *project* categories, adjusting the cost factor to account for the unbalanced proportion of negative instances to positive ones, and performing *leave-one-university-out* cross-validation for training and evaluation.

4. EVALUATION

We would like our evaluation to answer three questions: 1) what is the performance of a URL-only based system, 2) what is the performance of a non-source document system (i.e., anchor text plus URL features) and 3) whether URL features can improve the performance of source document systems. To answer these questions, we test different configurations of the system using the following features:

- **(U)RL text** - the web page's URL fragments. This is really four different feature sets generated by the techniques used for segmentation / expansion of the URL. U_b, U_r, U_i and U_f correspond respectively to the baseline, refined, information content reduction and title-based FST splitting algorithms.
- **(A)nchor text** - all tokens of the text contained in anchors pointing to the web page.
- **(T)itle text** - all tokens in the web page's <title> tag.
- **Page Te(X)t** - all tokens of the source document's text body (excluding tags in the <head>).

Separate features and counts are maintained for any token that appears in different feature sets (e.g., "cs" appearing in the title as well as in the URL), allowing us to independently assess the impact of each feature set. Performance is measured using the F_1 measure, which is defined as the harmonic mean of the precision and recall of the classifier. Table 2 shows the F_1 values for each SVM configuration for the four classes, macro-averaged over all four universities.

About 73% (3,078 pages) in the ILP dataset belong to the default *other* category. A majority-class classifier categorizes all pages as *other*, receiving 0 F_1 for all four categories. The performance of the inductive learner FOIL-PILFS (FP) as reported in [3], is a fairer comparison. This system used both the page text and anchor text as features, although formulated with frequency cutoffs.

Configuration (# of pages)	Course (245)	Faculty (153)	Project (84)	Student (558)	Macro Avg
U_b	13.5	23.4	35.6	15.8	22.1
U_r	50.2	24.2	36.4	16.6	31.9
U_i	50.2	31.8	35.0	15.7	33.2
U_f	52.7	31.5	36.3	15.6	34.0
A	41.7	21.3	32.9	16.9	30.9
T	39.2	12.6	54.8	16.0	30.6
X	55.1	41.6	61.3	12.2	42.5
A + U_i	43.1	30.6	37.2	20.1	32.7
A + U_f	43.4	27.4	37.0	20.9	32.0
X + A + T	59.3	40.3	67.2	26.2	48.3
X + A + T + U_i	60.8	40.0	66.5	25.3	48.2
X + A + T + U_f	60.9	40.9	66.5	25.3	48.4
FP (ϵ = .05)	53.1	43.4	21.7	54.6	43.2

Table 2: F_1 for different SVM configurations on the four page classification task using the ILP 98 WebKB dataset.

We make the following observations about the performance of the systems from these experiments on this task:

- Appropriate use of URL alone proves about three-fourths as effective using the page text itself and exceeds the performance of systems using the page title or its anchors words.
- The more complex segmentation / expansion methodologies do impact the performance of the system, but less so. We find that IC reduction and the FST expansion make changes in the original U_r segmentation in only 12.7% and 18.6% of the cases, respectively. As these methods only affect a small percentage of the URLs (unlike the change from the baseline to the refined URL parsing), their power to enhance performance is similarly limiting.
- Using both of non-source document techniques (i.e., anchor text and URLs) fails to improve results further than the best URL only based approach. Unfortunately as well, the URL only features also fail to improve the performance of knowledge-rich classifiers that have access to all available features.

5. CONCLUSION

We have quantified the performance of web page classification using only the URL feature. This ubiquitous feature of web pages, when treated correctly, exceeds the performance of some source-document based features. The effect of anchor text, a recent topic for IR research, underperforms the URL-based systems, even though the ILP dataset is highly interconnected (10,945 interlinks for the 4K pages). We expect the gap to widen as 1) most web classification problems have less interconnection, and 2) most general websites have more of a need for word segmentation methods (e.g., www.countrymusicfile.com). These results have encouraged us to assess the scalability of our technique on a larger problem based on Open Directory Project (ODP) categories as classification targets.

6. REFERENCES

[1] D. Bergmark. Collection synthesis. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, Portland, Oregon, USA, 2002.

[2] K. T. Lua and G. W. Gan. An application of information theory in chinese word segmentation. *Computer Processing of Chinese and Oriental Languages*, 8(1):115–124, 1994.

[3] S. Slattery and M. Craven. Combining statistical and relational methods for learning in hypertext domains. In *8th Int'l Conf. on Inductive Logic Programming*, 1998.

[4] A. Sun, E.-P. Lim, and W.-K. Ng. Web classification using support vector machine. In *4th Int'l Workshop on Web Information and Data Management (WIDM 2002)*, Virginia, USA, November 2002.