Fact or fiction: Content classification for digital libraries

Aidan Finn Nicholas Kushmerick Barry Smyth
Smart Media Institute, Department of Computer Science, University College Dublin
{aidan.finn, nick, barry.smyth}@ucd.ie

Abstract

The World-Wide Web (WWW) is a vast repository of information, much of which is valuable but very often hidden to the user. The anarchic nature of the WWW presents unique challenges when it comes to information extraction and categorization. We view the WWW as a valuable resource for the gathering of information for Digital Libraries. In this paper we will describe the process of extracting and classifying information from the WWW for the purpose of integrating it into digital libraries. Our efforts focus on ways to automatically classify news articles according to whether they present opinions or reported facts. We describe and evaluate a system in development that automatically classifies and recommends Web news articles from sports and politics domains.

1 Introduction

The WWW is a vast resource of digital information. It has the potential to be a potent information source for digital libraries. However, alongside this great potential is a significant problem: the information overload problem. It is becoming more and more difficult for the right users to locate the right information at the right time, and this is limiting the potential of the Web.

The ability to automatically retrieve and categorize documents from the WWW is an important step forward in helping to structure information on the Web in order to release its full potential - a valuable asset for any digital library system. However, retrieving and classifying documents from the Web presents a unique set of challenges. Information on the WWW may be unstructured, transient or unreliable and traditional methods often prove inadequate. We believe that new approaches are required [5].

The transient nature of information on the Web is not characteristic of other sources of information for digital libraries. Information often disappears from the web as it becomes obsolete (e.g. out of date news postings). New information may be extremely current but short-lived (breaking news, for example). The ability to quickly retrieve current topical information and channel it towards the appropriate users is an important service in any digital library system, and a valuable service for digital library users.

In this paper we outline such a system. In particular, we discuss the issues involved in automatically classifying Web articles, focusing on one particular classification problem: the classification of articles as either factual or opinionated. We detail the construction of our classifier, which includes a text extraction component for automatically identifying the body text of a Web article. Finally we evaluate the accuracy of the resulting classification system along with its ability to generalize well to unseen domains.

There is a large body of related work in information retrieval and automatic classification. We mention some samples from the area. Wrapper induction [4] is a technique for automatically generating wrappers for extracting information from web pages. Learning to classify documents as part of an ontology is discussed in [2]. The shopbot system [3] applies machine learning techniques to the problem of automatically extracting information from the web. A system for automatically classifying web sites into industry categories is described in [8]. The CORA project [7] uses machine learning techniques to automatically construct domain specific search engines. Subjectivity classification at the sentence level is discussed in [10]. Naïve Bayes is described in [6]. C4.5 is described in detail in [9].

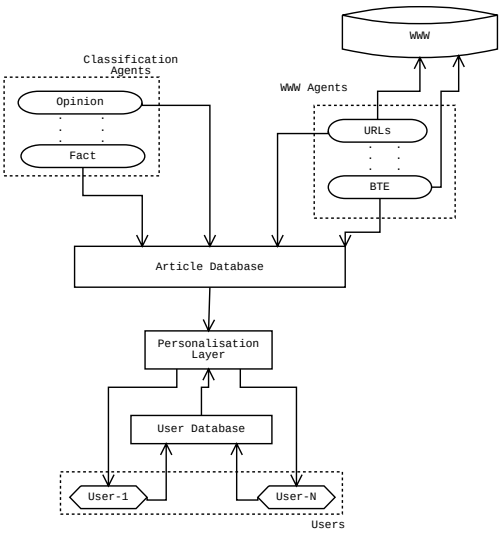

Figure 1: System Overview

2 Research Issues

The main aims of our research are to:

- promptly retrieve and classify new articles appearing on the web;
- develop a classification system that scales gracefully and transfers easily to new domains;
- classify articles according to varied and unusual criteria. E.g. to be able to recognize the tone of articles. Articles that present opinions are classified differently than factual articles; and
- enable personalization by retrieving documents from classes that are interesting to the user.

Fig. 1 presents an architecture that we propose will meet our objectives. We are instantiating this architecture in the HYPPIA system, a WWW service that monitors web sites for new articles, classifies these new articles using a number of text classifiers, and creates personalized digital libraries from this content. New classifiers can be added as the system grows. Classifiers should be built using as few domain specific features as possible in order to facilitate domain transfer. This system is currently in development; in this paper we describe our preliminary results.

The core of the system is the article database. This acts as a central information repository for all parts of the system. A set of WWW agents retrieve information from the web and store it in this database. A URL agent watches web sites for new articles and stores their URLs. The BTE agent (described in Sec. 3) extracts the main body of text from new articles for use by our text classification algorithm (described in Sec. 4).

A set of classification agents classify articles in the database. Each classifier is responsible for a particular class. Classifiers are built by hand but we are currently exploring techniques for automating this process. Currently, we have incorporated two opinion classifiers. It is anticipated that more classifiers will be added in the future.

In the remainder of this paper we focus on two elements of this system: automatic extraction of body text, and subjectivity classification of documents on the WWW. The opinion classifier recognizes articles that express the opinion of the author. Articles of this kind often take the form of columns or editorials. These articles contrast with those which just report facts. We focus on this form of classification because different users may have different requirements or preferences as to which type of article they prefer. When developing the opinion classifier, our motivation was to exclude domain specific features in order to easily migrate the classifier to new domains. For example, if we train our system in the domain of football news, we want it to generalize appropriately when tested with politics articles.

Figure 2: A typical news article.

3 Body Text Extraction

Articles published on the WWW often contain extraneous clutter. Most articles consist of a main body which constitutes the relevant part of the particular page. Surrounding this body is irrelevant information such as copyright notices, advertising, links to sponsors, etc. By identifying this main body of the article and basing our classifiers on features that occur within this area we anticipate improving the accuracy of classifiers. For example, images and links occurring within this body are likely to be very relevant to the article in question, while images and links outside this body are far less likely to provide interesting information about the article. Similarly when extracting text from a web page for use in classification, text extracted from the article body is much less likely to mislead the classifier than text extracted from the surrounding clutter. Fig. 2 shows a typical news article with the main body of text marked. Below the body is a generic copyright notice, while on each side are links to advertisers and unrelated articles.

Automatically identifying the region of the web page that contains the main body of text is useful when building a classifier. We found that text from outside the body of an article was less likely to be relevant to the document and therefore more likely to mislead a text classifier.

Identifying the main body of a web page in a general robust manner is a difficult information extraction problem. Our approach is to view a web page as as consisting of two kinds of tokens: HTML tag tokens and text tokens (i.e., words). Thus an HTML page can be represented as a sequence of bits B, with a $B_n = 0$ indicating that the n'th token is a word, and $B_n = 1$ indicating a tag. Fig. 3 shows a graph of the cumulative distribution of tags, as a function of the position in the document, for the article from Fig. 2. Our extraction algorithm is based on the idea that the ends of the "plateau" in the middle of the distribution correspond to the points at which the main body of the article begins and ends.

The problem can now be viewed as an optimization problem. We must identify points i and j such that we maximize the number of tag tokens below i and above j, while simultaneously maximizing the number of text tokens between i and j. The text is only extracted between i and j. We chose a simple objective function $T_{i,j}$ where:

$$T_{i,j} = \sum_{n=0}^{i-1} B_n + \sum_{n=i}^{j} (1 - B_n) + \sum_{n=j+1}^{N-1} B_n$$

The advantage of this method of text extraction is that it does not require any parameters, and can extract text from different web sites without the need for site specific wrappers.

To summarize, we use this text extraction algorithm to extract the body of text from web pages[1] for use

[1] This algorithm is available as a perl module from the project web-page at http://smi.ucd.ie

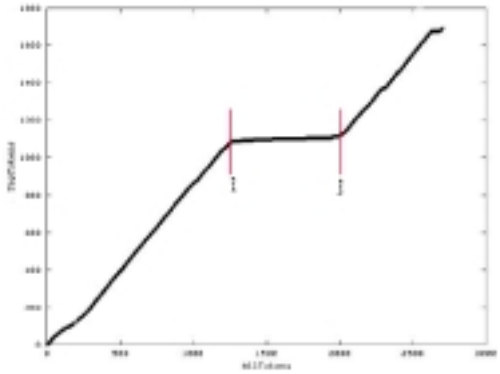

Figure 3: The cumulative distribution of tags, as a function of the position in the document. The central "plateau" corresponds to the body of the text highlighted in Fig. 2.

by our classifiers. This was motivated by the observation that text from outside the main body often misled the classifiers.

4 Opinion Classification

The classifiers built so far are based on the text extracted from the body of the articles. The system operates by watching news sites for new articles and classifying them. Articles that appear on those sites are often either news articles, which just report facts, or articles that present the opinion of the author. A good example of the former are articles from the Reuters news service, while the latter often take the form of columns or editorials. We investigate the possibility of automatically recognizing these opinionated articles.

In building our opinion classifier, one of our main aims was to have high domain transfer. Thus we did not want to classify based on features unique to the football domain, because then it would not generalize if tested on a new domain, such as politics.

As a baseline, we first considered a classifier that used the occurrence of words in the text as features, in conjunction with a naïve Bayes classifier. It was anticipated that using words that occurred in the text as features would not produce a general classifier. We expected that this approach would not transfer well to new domains or perform well on articles from sites it had never seen before. In fact, it was experiments with this approach that led us to develop our method for body text extraction. We found that using all text from a web page would either mislead the classifier or provide predictive features which were too strong. For example if one particular site in the training data produces only opinion articles, something such as a copyright notice in this page will provide a very strong feature for classification. While this feature may be useful in a single domain system, it will not be very useful if for example the classifier is to be used to classify articles on politics.

In contrast to this baseline, our second approach examines the type of language in the document. Intuitively, we expect that the kind of language used in opinion documents is different, perhaps more expressive than that used in factual articles. For example, the number and type of adjectives in a sentence is indicative of its subjectivity [10]. In using elements of the language style as features, rather than language content we hope to produce a classifier capable of generalizing well to unseen domains. In particular, we first processed the documents using Brill's Parts-Of-Speech (POS) tagger [1], and then represented a document as the the fraction of words for each POS. We then used C4.5 to learn a decision tree based on these POS features. While a hybrid approach using both word and lexical features is probably the best overall solution, in these preliminary experiments our classifier relies just on the POS features.

5 Evaluation

Experiments used documents from two domains, football and politics, in order to evaluate accuracy as well as domain transfer. A corpus of documents was spidered from the web and then manually classified as being

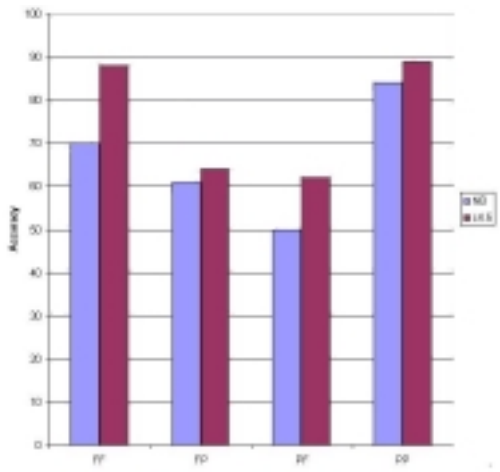

Figure 4: Opinion classifier Accuracy

if %unique>0.65 and %superlativeAdverb<=0.001
⟹ class notOpinion
if %comparativeAdjective<=0 and %wh-Adverb<=0.01
⟹ class notOpinion

Table 1: Sample Rules

opinion or non-opinion. Our corpus comprises 350 football articles (of which 174 are classified as opinion based) and 230 politics articles (of which 155 are classified opinion based). Our experiments use ten-fold cross validation, with a 80%/20% training/testing split.

As described above, we compared two classifiers. The first uses occurrences of words as features in conjunction with a naïve Bayes classifier. The second uses POS statistics in conjunction with a C4.5 classifier.

Our results are shown in Fig. 4. The vertical axis shows classification accuracy, defined as the percentage of articles in the test set that were classified correctly. The horizontal axis indicates the training and test data. For example, FP indicates the classifier was trained on football documents and tested on politics documents. Results are shown for both the naïve Bayes and C4.5 classifiers. Some sample rules generated by the C4.5 learning algorithm are shown in table 1.

Results show that the classifier based on POS statistics (C4.5) performs better than the classifier based on word occurrence statistics (NB) in all cases. We conclude that the kind of language used in a document is a better indication of subjectivity than the content of the document.

In cases where the classifier was trained and tested in the same domain (FF and PP), the accuracy of the C4.5 classifier is quite impressive (86% and 88% respectively). In the football domain (FF) this is significantly better than the NB approach although NB performs closer to C4.5 in the politics domain(PP). This may indicate that the relevance of content to subjectivity varies with different domains.

We expected that the classifier based on POS statistics would generalize better across domains. This was tested by training a classifier in one domain and testing it in the other domain. In the PF case, the C4.5 classifier performs significantly better than the naïve Bayes classifier. However, in the FP case, C4.5 performs only slightly better than naïve Bayes. We conclude that the POS approach is better suited to generalizing to new domains compared to the naïve Bayes algorithm.

Note also that the C4.5 results are more stable than the NB results. Both cases that were trained and tested in the same domain return approximately the same accuracy. Similarly, both cases that were trained and tested in different domains return approximately the same result. This may indicate that while the content of a document may be of varying relevance to a subjectivity classifier, the type of language used is a more reliable indication of subjectivity.

6 Conclusion

Automatic document classification is an important way of bringing order to the Web and is a key enabling technology for the next generation of digital library systems. In this paper we have focused on one particular classification problem dealing with recognizing the nature of Web articles in terms of whether they represent reported facts or author's opinions.

Experimental results with our classifier indicate that it is possible to accurately classify documents according to whether they represent facts or opinions. These results demonstrate that it is possible to achieve high levels of accuracy by using just the Parts-Of-Speech statistics as features. Our experiments bear out the intuition that the kind of language used in a document is a better indication of subjectivity than the actual content of the document.

Currently we have implemented a subjectivity classification agent. In future, more classification agents will be added to the system. For example, in the football domain, it would be useful to have classifiers for each of the major football clubs. These classifiers could potentially be constructed automatically as there are already many sources of pre-classified training data on the WWW(e.g. each teams home page). We are implementing a demonstration system that allows users to provide feedback on the accuracy of the classification agents to facilitate further experiments.

Our ultimate goal is personalized digital libraries. Our approach is that, initially, users could be presented with articles from the classes they are interested in. Over time, the system could build a classifier for the user, as their interests become known from the articles that they read.

Acknowledgements

This research was funded in part by grant N00014-00-1-0021 from the US Office of Naval Research.

References

[1] Eric Brill. Some advances in transformation-based parts of speech tagging. In *AAAI-94*.

[2] Mark Craven, Dan DiPasquo, Dayne Freitag, and Andrew McCallum. Learning to extract symbolic knowledge from the world wide web. In *AAAI-98*.

[3] Robwert D. Doorenbos, Oren Etzioni, and Daniel S. Weld. A scalable comparasion-shopping agent for the world-wide web. In *ACM Agents '97*.

[4] Nicholas Kushmerick, Daniel S. Weld, and Robert Doorenbos. Wrapper induction for information extraction. In *IJCAI-97*.

[5] Sean Luke, Lee Spector, David Rager, and James Hendler. Ontology-based web agents. In *AA-97*.

[6] Andrew McCallum and K. Nigam. A comparision of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*.

[7] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. A machine learning approach to building domain-specific search engines. In *IJCAI-99*.

[8] John M. Pierre. On the automated classification of web sites. In *Linkoping Electronic Articles in Computer and Information Science*, volume 6. 2001.

[9] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.

[10] Janyce M. Wiebe. Learning subjective adjectives from corpora. In *AAAI-00*.