Classifying Web content for a corporate digital library

Ian Thurlow

BT

Abstract: The integration of relevant Web content into corporate digital libraries is expected to be of significant benefit when it is classified into existing classification schemes and made accessible through the library's search and browse tools. The integration of Web content into a digital library, however, raises some concerns with regard to the quality and classification of that content. The accurate automatic classification of Web content into subject categories is fundamental to Web content integration into a digital library, particularly if the metadata generated by an automatic classifier is used to create indexing elements on which the search and browse tools will operate. Ideally, the quality of the classifications given by an automatic classifier should approach that of a level that is consistent with the classifications found in bibliographic record databases. This paper gives an overview of some research work that is examining the use of a controlled vocabulary to classify previously unseen Web content. A prototype classifier has been developed as part of this work. A preliminary inspection of the subject categories presented by the classifier suggests that it can produce sound classifications of Web content against an existing classification scheme. Further work is required to establish the effect that each parameter of the classifier has on the quality of the classification.

1. Introduction

A great deal of valuable, high-quality information resides in corporate digital libraries. Likewise, the Web holds an abundance of useful information, much of which could be of far greater value if integrated with existing digital library content and made accessible through the library's search and browse tools. The Semantic Web offers the prospect of a new approach to the integration and management of information on the Web [1], the fundamental principle of which is to derive, create and make use of machine-processable metadata in the form of semantic annotations [2]. The incorporation of semantic web technology into knowledge tools and applications is expected bring a number of benefits to the end user, in particular such applications are expected to help people gain access to relevant information more efficiently and more effectively [3]. This is particularly true for search and browse applications. The huge potential of semantically enabled applications will, however, only be realised when the semantics of the content can be extracted with a sufficient level of accuracy, enabling the search tools to exploit those semantic annotations to their best effect. High quality annotations will be central to the success of the technology.

2. Classification of web content into digital libraries

The Semantically Enabled Digital Technologies (SEKT) project [4] is exploring the use of semantic web technology to improve information access. The BT digital library is being used as a SEKT technology case study setting [5][6]. Despite many of the advantages that semantic technology is expected to bring to digital libraries [7], the integration of Web content raises some concerns, particularly with regard to the volume of content, the quality of that content, the level of trust in the source, and the accuracy of its classification. Even if the quality of the content and level of trust in the source could be assured, the sheer volume of information on the Web will make it impractical, in terms of effort and cost, to classify Web content into the library using manual annotation methods. Consequently, automatic classification techniques [8] will need to be deployed. In order to integrate Web content into BT's digital library, and make it more accessible to library tools and applications, it will be necessary to classify that content against the current classification schemes, and to associate the metadata relating to the classification with that content (the assignment of one or more pre-defined category labels to natural texts is discussed in [9]).

The accurate classification of Web content is considered to be fundamental to its integration into the digital library, as the metadata generated by the automatic classifiers is likely to be used by the indexing components to create indexing elements on which the search and browse tools will operate. The importance of the quality of any automatic classification must not be underestimated; if the quality of the metadata on which the tools operate is poor, the operation of those tools in return is likely to be poor. In the setting of a digital library, it is anticipated that previously unseen Web content will need to be classified automatically with a level of quality that approaches that is which is normally found in bibliographic record databases.

3. Classification based on controlled language term co-occurrence

BT subscribes to approximately 1000 on-line publications, which gives end-users access to the full-text of over 900,000 scientific and business articles and papers. In addition, BT procures the Inspec and ABI bibliographic record databases, giving access to over 4 million bibliographic records. The format for each bibliographic record is based on ISO 2709 (which is based on the Library of Congress MARC format [10]). Each record comprises a number of fields, e.g. the record for a journal article would include the full title of the article, the names of the authors, a publication date, an abstract, a number of classification codes and a set of controlled indexing terms. An excerpt from a sample Inspec bibliographic record is shown in Appendix A (figure 1). In the case of the Inspec records, the controlled indexing terms are taken from the Inspec Thesaurus [11]. The controlled indexing terms for the ABI records are taken from the ProQuest Controlled Vocabulary of Subject Terms [12]. Human indexers assign the controlled indexing terms in each record.

A prototype classifier has been developed, which assigns subject categories from the ABI and Inspec controlled vocabularies to content retrieved from the Web. The classifier identifies the occurrence of controlled indexing terms in the text, and selects those terms for classification that are deemed most significant. The significance of each controlled indexing term is not only dependent on its frequency of occurrence in the text, but also on its inter-dependencies with other controlled indexing terms. The inter-dependencies of the controlled indexing terms are calculated from the sets of bibliographic records prior to any classification.

Assume that *N* unique controlled indexing terms are identified in a set of bibliographic records, e.g. Inspec.

For each of the *N* controlled indexing terms in that set of records, a vector of related controlled indexing terms is constructed. More precisely, for each controlled indexing term *y*, a vector

$$\left\{a_y^i\right\} = \left\{a_1^y, a_2^y, ...a_N^y\right\}$$

is constructed, where a_i^y is a value that is dependent on frequency of co-occurrence of the controlled indexing term *y* and the *i*th controlled indexing term (a_i^y is zero if the controlled indexing term *y* does not co-occur in any record with the *i*th controlled indexing term).

For each document *D* to be classified, a vector

$$\left\{x_i^D\right\} = \left\{x_1^D, x_2^D, ...x_N^D\right\}$$

is constructed, where:

$$x_i^D = \begin{cases} 1 & \text{If the } i\text{th controlled indexing term occurs in } D\text{, where } i = 1, \ldots, \text{N} \\ 0 & \text{otherwise} \end{cases}$$

The ranking value RV_D^y for the controlled indexing term *y* is calculated from the dot product of the vector $\left\{a_y^i\right\}$and the vector $\left\{x_i^d\right\}$

$$RV_D^y = \sum_{i=1}^{N} x_i^D \bullet a_i^j$$

The highest ranking controlled indexing term *y* is the term that maximises RV_D^y , i.e.

$$arg\left(Max_y RV_D^y\right)$$

The co-occurrence vectors are used in this way to rank order controlled indexing terms to a previously unseen Web page and the highest ranking terms indicate the classification.

4. Results to date

A number of Web pages were collected from a small number of technology-oriented web sites. Each web site was selected for its trustworthiness and the quality of its content. The text from each web page was extracted and then classified. An example of an article taken from the BBC web site along with its classification is shown in Appendix A (figure 2 and figure 3). Although an inspection of the classifications presented by the classifier for a number of web pages suggests that it is capable of producing sensible classifications for previously unseen Web content, it must be emphasised that the quality of those classifications has not been assessed formally.

5. Evaluation

A formal evaluation of the classifier will be completed in the SEKT case study. The following evaluation method will be used. A set of bibliographic records covering a specific topic will be selected. This set of records will be divided into a training set and a test set. A set of controlled indexing term co-occurrence vectors will be generated from the controlled index field of each record in the training set. The full text of each article that is associated with the bibliographic records in the test set will then be classified by the classifier. The set of classifications generated by the classifier will be compared to the set of classifications originally assigned to that article in its associated bibliographic record.

6. Further work

Future work will concentrate on establishing the effect of each of the classifier's parameters on the quality of the classification and evaluating the approach on a greater range of documents. For example, many terms from the controlled vocabulary comprise more than a single word, e.g. 'control systems' or 'control system analysis'. These noun phrases are likely to more useful as a discriminator for the purpose of classifying a text, e.g. if the controlled language term 'knowledge representation languages' was identified in the text, this is likely to be more significant than the phrases 'knowledge representation' and 'languages' occurring in a text separately. The effect on the classification of matching deconstructed elements of controlled vocabulary multi-word terms will be investigated.

A number of problems are also anticipated in relation to mapping the full-text of a document into the precise language of the controlled vocabulary. Short documents, in particular, are expected to be problematical due to the sparseness of controlled indexing terms within the text. Furthermore, the ambiguity present in natural language text is expected to cause problems. For example, word polysemy, where the same word can be used to express multiple meanings, does not arise when classifying a text manually against the controlled vocabulary of a thesaurus as the person classifying the document will understand the sense of the word from its context. The same does not necessarily apply to automatic classifiers, in that a controlled vocabulary term could have alternative meanings in different texts, e.g. the word *interrupts* can be used to mean interject, or can be used in the context of a set of signals that get the attention of a computer's CPU. This gives the potential for the classifier to misinterpret the meaning of a controlled vocabulary term that occurs in a text if the decision to classify that text against the controlled term is based purely on the frequency of occurrence of that word alone. As some controlled indexing terms could have alternative meanings in a text, the sense of a word needs to be established. It is anticipated that the co-occurrence vectors will help establish a particular sense of a word.

7. Conclusions

A prototype classifier has been developed that uses vectors of co-occurring controlled indexing terms to rank order the subject classifications given to a previously unseen web page. Although the approach has not been evaluated on a statistically significant range of documents, a preliminary inspection of the classifications suggested by the classifier shows that it is capable of producing sound classifications for previously unseen content. A detailed evaluation of the classifier will be undertaken as part of the ongoing research work in the SEKT project.

8. References

[1] Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American,* May 2001.

[2] Davies, J., Fensel, D., and van Harmelen, F (Editors). *Towards the Semantic Web: Ontology-Driven Knowledge Management*. England: John Wiley & Sons Ltd, 2003. ISBN: 0-470-84867-7.

[3] Davies, J., Duke, A., Kings, N., Mladenic, D., Bontcheva, K., Grcar, M., Benjamins, R., Contreras, J., Blazquez Civico, M., and Glover, T. Next generation knowledge access. Journal of Knowledge Management, Volume 9, Number 5, 2005. ISBN 1-84544-805-7. ISSN 1367-3270.

[4] http://www.sekt-project.com/

[5] Warren, P., and Alsmeyer, D. Applying semantic technology to a digital library: a case study. Library Management, May 2005, Volume: 26 Issue: 4/5 Page: 196 – 205, ISBN 1-84544-141-9. ISSN: 0143-5124.

[6] Davies, J., Studer, R., and Warren, P. (Editors). Semantic Web Technologies: Trends and Research in Ontology-based Systems. England: John Wiley & Sons Ltd, 2006. ISBN: 0-470-02596-4.

[7] Sure, Y., and Studer, R. Semantic Web Technologies for Digital Libraries. Library Management, May 2005, Volume: 26 Issue: 4/5 Page: 190-195, ISBN 1-84544-141-9. ISSN: 0143-5124.

[8] Sebastiani, F. Machine Learning in Automated Text Categorzation, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp 1-47.

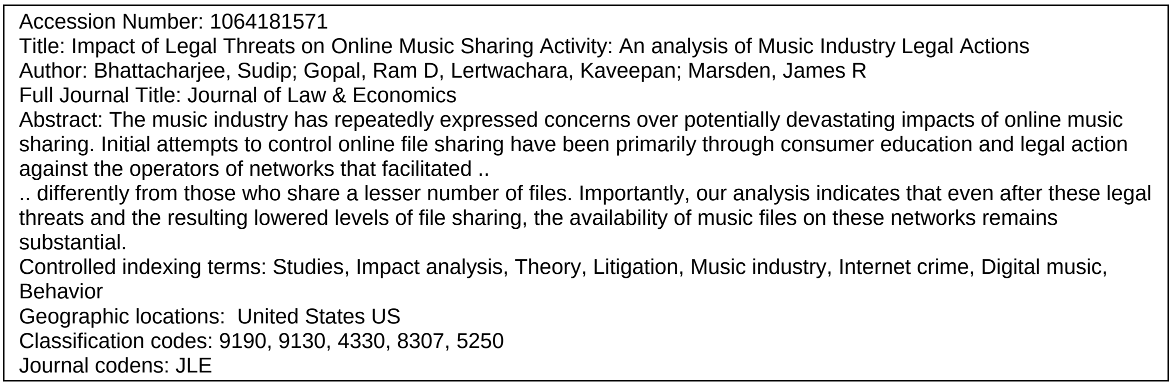
[9] Dumais, S. Data-driven approaches to information access. Cognitive Science 27 (2003) 491-524.

[10] http://www.loc.gov/marc/

[11] http://www.iee.org/publish/support/inspec/document/thes/

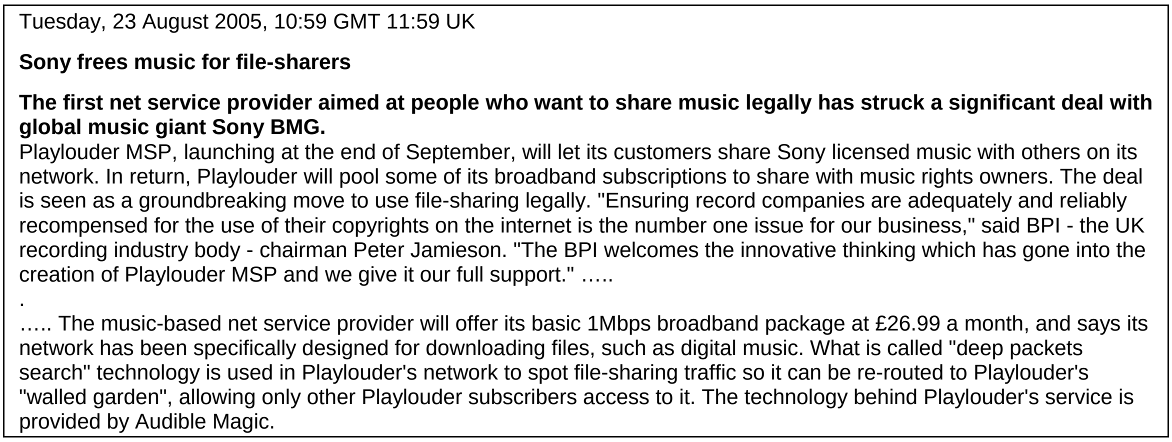
[12] http://www.chadwyck.com/products_pq/controlled-vocab/

Appendix A: Bibliographic record and classification of a Web article



Accession Number: 1064181571
Title: Impact of Legal Threats on Online Music Sharing Activity: An analysis of Music Industry Legal Actions
Author: Bhattacharjee, Sudip; Gopal, Ram D, Lertwachara, Kaveepan; Marsden, James R
Full Journal Title: Journal of Law & Economics
Abstract: The music industry has repeatedly expressed concerns over potentially devastating impacts of online music sharing. Initial attempts to control online file sharing have been primarily through consumer education and legal action against the operators of networks that facilitated ..
.. differently from those who share a lesser number of files. Importantly, our analysis indicates that even after these legal threats and the resulting lowered levels of file sharing, the availability of music files on these networks remains substantial.
Controlled indexing terms: Studies, Impact analysis, Theory, Litigation, Music industry, Internet crime, Digital music, Behavior
Geographic locations: United States US
Classification codes: 9190, 9130, 4330, 8307, 5250
Journal codens: JLE

Figure 1. An excerpt of an example bibliographic record

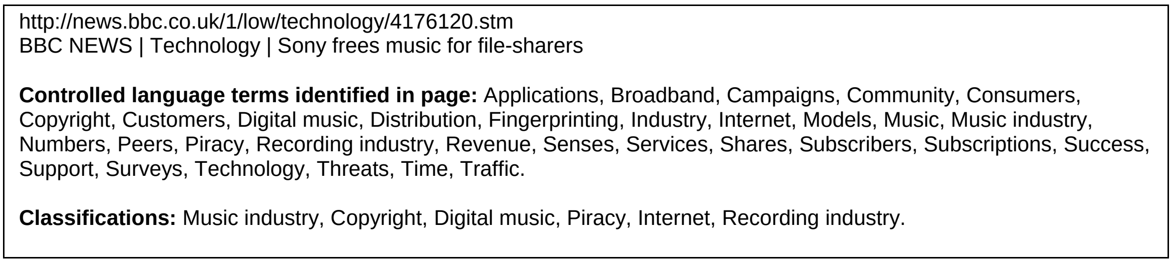


Tuesday, 23 August 2005, 10:59 GMT 11:59 UK

Sony frees music for file-sharers

The first net service provider aimed at people who want to share music legally has struck a significant deal with global music giant Sony BMG.
Playlouder MSP, launching at the end of September, will let its customers share Sony licensed music with others on its network. In return, Playlouder will pool some of its broadband subscriptions to share with music rights owners. The deal is seen as a groundbreaking move to use file-sharing legally. "Ensuring record companies are adequately and reliably recompensed for the use of their copyrights on the internet is the number one issue for our business," said BPI - the UK recording industry body - chairman Peter Jamieson. "The BPI welcomes the innovative thinking which has gone into the creation of Playlouder MSP and we give it our full support." …..
.
….. The music-based net service provider will offer its basic 1Mbps broadband package at £26.99 a month, and says its network has been specifically designed for downloading files, such as digital music. What is called "deep packets search" technology is used in Playlouder's network to spot file-sharing traffic so it can be re-routed to Playlouder's "walled garden", allowing only other Playlouder subscribers access to it. The technology behind Playlouder's service is provided by Audible Magic.

Figure 2. An excerpt of an example web article.



http://news.bbc.co.uk/1/low/technology/4176120.stm
BBC NEWS | Technology | Sony frees music for file-sharers

Controlled language terms identified in page: Applications, Broadband, Campaigns, Community, Consumers, Copyright, Customers, Digital music, Distribution, Fingerprinting, Industry, Internet, Models, Music, Music industry, Numbers, Peers, Piracy, Recording industry, Revenue, Senses, Services, Shares, Subscribers, Subscriptions, Success, Support, Surveys, Technology, Threats, Time, Traffic.

Classifications: Music industry, Copyright, Digital music, Piracy, Internet, Recording industry.

Figure 3. Classification of the Web article.