Supporting Systematic Reviews using Text Mining

Sophia Ananiadou[1], Rob Procter[2], Brian Rea[1], Yutaka Sasaki[1] and James Thomas[3]

[1]National Centre for Text Mining, School of Computer Science, University of Manchester

[2]National Centre for e-Social Science, School of Social Sciences, University of Manchester

[3]Social Science Research Unit, Institute of Education, University of London

Email address of corresponding author: Sophia.Ananiadou@manchester.ac.uk

Abstract. In this paper, we describe how we are using text mining solutions to enhance the production of systematic reviews. This collaborative project also serves as a proof of concept and as a testbed for deriving requirements for the development of more generally applicable text mining tools and services.

Introduction

Like the natural sciences, the social sciences are facing a 'data deluge' (Hey and Trefethen, 2003) which exceeds the capacity of current research methods and tools. One example is the challenge faced in literature surveys ('systematic reviewing') by the rapid growth in the research literature. Another is the challenge posed by new sources of data such as the WWW (news and corporate sites, wikis, blogs, etc.), digital communications (email, newsgroups, speech, SMS) and transactional records (purchases, etc.) which offer an extremely rich resource for research. Equally, the emergence of research, learning and teaching repositories in recent years containing textual data sources and materials offers the opportunity to analyse across multiple data collections in different locations. The WWW archive, for example, currently contains 55 billion Web pages or 2 petabytes ($2x10^{15}$) of data and is growing at the rate of 20 terabytes ($20x10^{12}$) per month. If the social sciences are to deal with this data deluge, they must harness powerful new text mining technologies. In practical terms, this requires the development of a set of interoperable text mining tools and services which can be integrated into different research practices and user communities.

The ASSERT project

In this paper, we describe the progress of the ASSERT project[1] and how we are using text mining solutions to enhance the production of systematic reviews. ASSERT is led by the UK National Centre for Text Mining[2], in collaboration with the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)[3] (University College, London) and the

[1] http://www.nactem.ac.uk/assert/
[2] http://www.nactem.ac.uk/
[3] http://ioewebserver.ioe.ac.uk/ioe/

National Centre for e-Social Science (NCeSS)[4]. This project also serves as a proof of concept and a test bed for deriving requirements for the development of more generally applicable text mining tools and services for the social sciences. In this paper, we will discuss in more detail the application of text mining techniques for improving the search and screening strategy of systematic reviewing for the domain of *rehabilitation of people with mental health issues.*

Project design and development methodology

A range of methods have been devised over the past 20 years to tackle the challenge of identifying user requirements and usability issues of IT systems and securing the effective involvement of users of the life time of a project (see, for example, Jirotka and Goguen, 1994). Interviews, focus groups, workshops and prototyping all have their role to play. More recently, ethnography, with its focus on detailed observation of how work actually gets done – the circumstances, practices and activities that constitute the 'real, local' character of work, its working divisions of labour, expertise, patterns of communication, coordination and use artefacts – has been added to the repertoire (Anderson, 1994). The value of ethnography lies in recognition of the need to study in context precisely how work is done which, in turn, facilitates recognition and understanding of user requirements and IT systems usability issues.

We are making use of many of these techniques in an iterative and user-driven process of work place studies, requirements gathering, rapid prototyping, evaluation and refinement to ensure that user requirements are systematically identified and tracked over the course of the project. The key to our method, however, is to 'embed' the text mining developers within the users' work place. The aim is to foster collaborative working between text mining tool developers and users, and thereby facilitate the 'co-realisation' of the system (Hartswood et al., 2005; Hartswood et al., 2007). This approach is critical if we are to understand how to embed text mining services within established routines of research practice and resource use, and how these may evolve as users begin to apply new tools in their work.

An overview of systematic reviewing

Before undertaking any new policy, practice or research it is essential to find out what is already known about an issue in a fair and unbiased manner. This may include the findings of individual research studies that might, alone, be limited in their applicability and vulnerable to bias. In order to minimise this bias, a large number of people and organizations, such as the Cochrane Collaboration[5], CRD[6] and the EPPI-Centre have developed methods for locating research evidence and synthesizing it in order to inform decision-making. The EPPI-Centre has developed ways of conducting literature reviews in a systematic way, which provide users with a 'short-cut' to relevant evidence.

Currently, systematic reviewing is performed mostly manually and encounters many problems. Part of the problem is due to the proliferation of textual information which means that the quantity of potentially relevant literature retrieved in the early stages of a review can become unmanageable – and with the literature expanding by several thousand papers per week, it is obvious that no individual can read them all.

Reviewers have been accustomed to sacrificing specificity in searches in order to ensure they have not missed any relevant studies, leading to searches which yield large numbers of 'hits'.

[4] http://www.ncess.ac.uk/
[5] http://www.cochrane.org/
[6] http://www.york.ac.uk/inst/crd/report4.htm/

They then download the titles and abstracts and screen them manually. This is the most time-consuming part of the process and can involve tens of thousands of titles and abstracts. Complex systematic reviews can take more than a year to complete with up to half time being spent searching and screening hits. This is problematic because policy-makers and practitioners often need to know the state of research evidence over a much shorter timescale than current methods allow. It can lessen the likelihood that research evidence will be used at all, with consequential dangers for people affected by policies or practices developed in the absence of a firm evidence base (Chalmers, 2003).

Systematic reviews proceed through the following stages:

1. *Searching:* extensive searches are carried out in order to locate as much relevant research as possible according to a query. These searches include electronic databases, scanning references lists and searching for unpublished literature.

2. *Screening*: narrows the scope of search by reducing the collection to only the relevant documents to a specific review. The aim is to highlight key evidence and results that may impact on the policy.

3. *Synthesizing*: correlates evidence from a plethora of resources and summarises the results.

Applying text mining to systematic reviewing

Informed by the study of systematic reviewing practices and requirements gathering, text mining techniques are being used to support these stages as follows (see Figure 1):

1. Searching can be improved by using *query expansion* techniques based on the most important concepts (terms) (Frantzi, Ananiadou and Mima, 2000) similarities among terms but also ontologies and thesauri.

2. Screening can be improved by using *document clustering* which groups documents into topics. These topics-clusters ideally correspond to a topic that is shared by all the documents they contain and by no other document in the collection. Visualisation allows the reviewer to see the associations between documents. By selecting topics the user obtains an overview of the documents in the sub-collection and is able to browse visually for alternative categories. *Document classification* automatically assigns documents into existing categories, generating subsets of documents focused on a specific topic, allowing for more efficient and accurate analysis during subsequent stages of information filtering (Joachims, 1998; Sebastiani, 2002). Clustering identifies a set of clusters based upon the most significant subset of the documents which in turn provide the relevant categories to create a training set. A classifier is then automatically built for each cluster at runtime and used to place each uncategorised document into a category. Multi-topic classification is useful for systematic reviewing as single documents may be relevant to multiple review topics.

3. Synthesising can be improved by using an adaptable multi-document summarisation driven by user defined *viewpoints* (Bollegala, Okazaki and Ishizuka, 2006; Lin and Hovy, 2001). The selection of salient sentences for each viewpoint will be based on *chronological ordering*. For multi-document summarisation systems it is important to produce a coherent arrangement of the extracted sentences from multiple documents. Source documents for a summary may have been written by different authors, and have

different text styles, dates, etc therefore arranging the salient sentences in a coherent manner is important. We select sentences from each document based upon the significance of its terms which are combined with classification techniques to discover the most relevant passages within the important sections of a document such as *Introduction, background, methodology, results, conclusions.* This technique provides a more informative overview of the document than a traditional abstract.

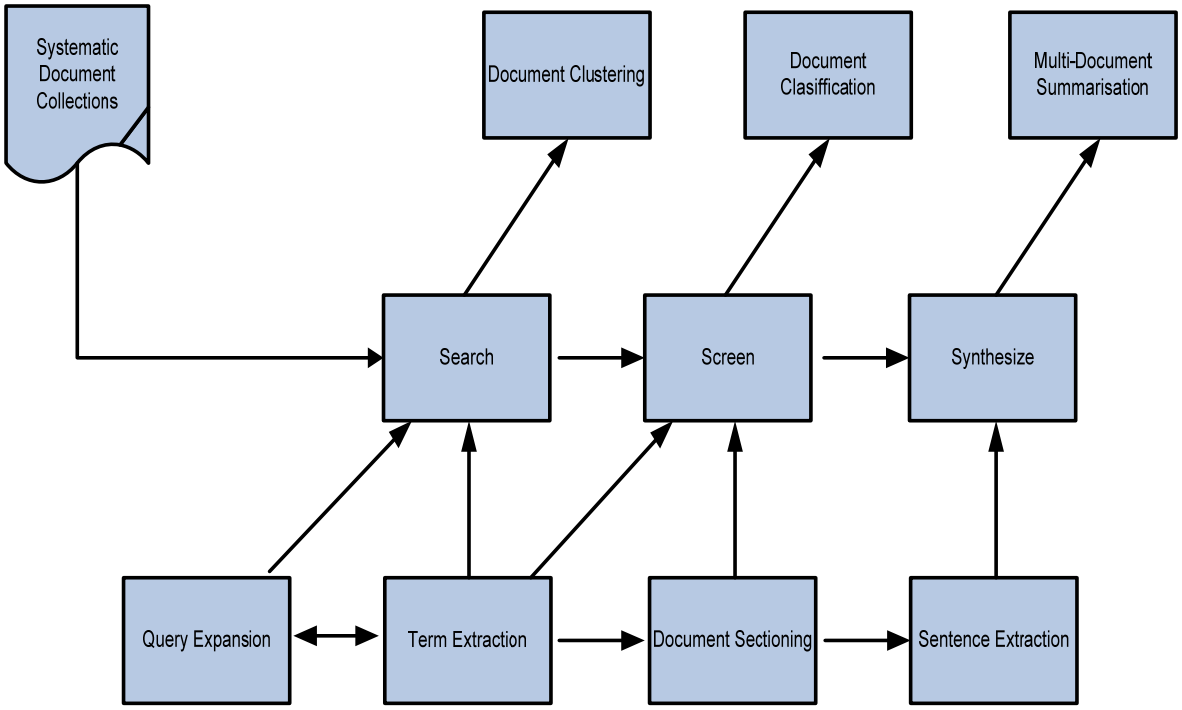

Figure 1: Text mining techniques and workflow for enhancing systematic reviewing

Improving the search strategy

Currently, searching in systematic reviews is performed manually. Reviewers are searching bibliographic databases based upon a defined search strategy i.e. an exhaustive list of manually constructed keywords by EPPI reviewers that detail the issues important to the search. Reviewers use sets of *inclusion* and *exclusion* criteria to determine if any given article is relevant to the review. Text mining improves the search strategy by using an associative search which discovers the set of documents most similar to a given document. Associative search is based on the assumption that documents sharing similar words mention similar topics. In our system, we place more emphasis on the significant words (terms) in a collection of documents. We first extract the most significant words in a collection of documents by using NaCTeM's TerMine service.[7] TerMine extracts and automatically ranks technical terms based on our hybrid term extraction technique C-value (Frantzi, Ananiadou and Mima, 2000). The C-value scores are combined with the indexing capabilities of Lucene 2.2[8] for full text

[7] http://www.nactem.ac.uk/software/termine/
[8] http://lucene.apache.org/index.html

indexing and searching. C-value scores are statistical measures used to evaluate how important a term is to a document or to a collection of documents.[9]

The output of associative searching is a ranked list of documents **similar** to the original document. Associative search is used and expanded to identify subsets of closely related documents (document clustering). Document similarities are calculated based upon the term/document vectors, which are weighted based on tf*idf. We use the **lingo** algorithm (Osiński and Weiss, 2005a) for document clustering. Clustering is based on finding a set of representative terms (that are not too similar) and their associated features. Clustering works better with larger document collections as this can reduce a lot of the noise and allows for a more complete view of the domain. This algorithm automatically generates human-readable descriptive labels for each of the clusters, allowing the reviewer to gain a quick overview of the collection based upon the variation of the labels.

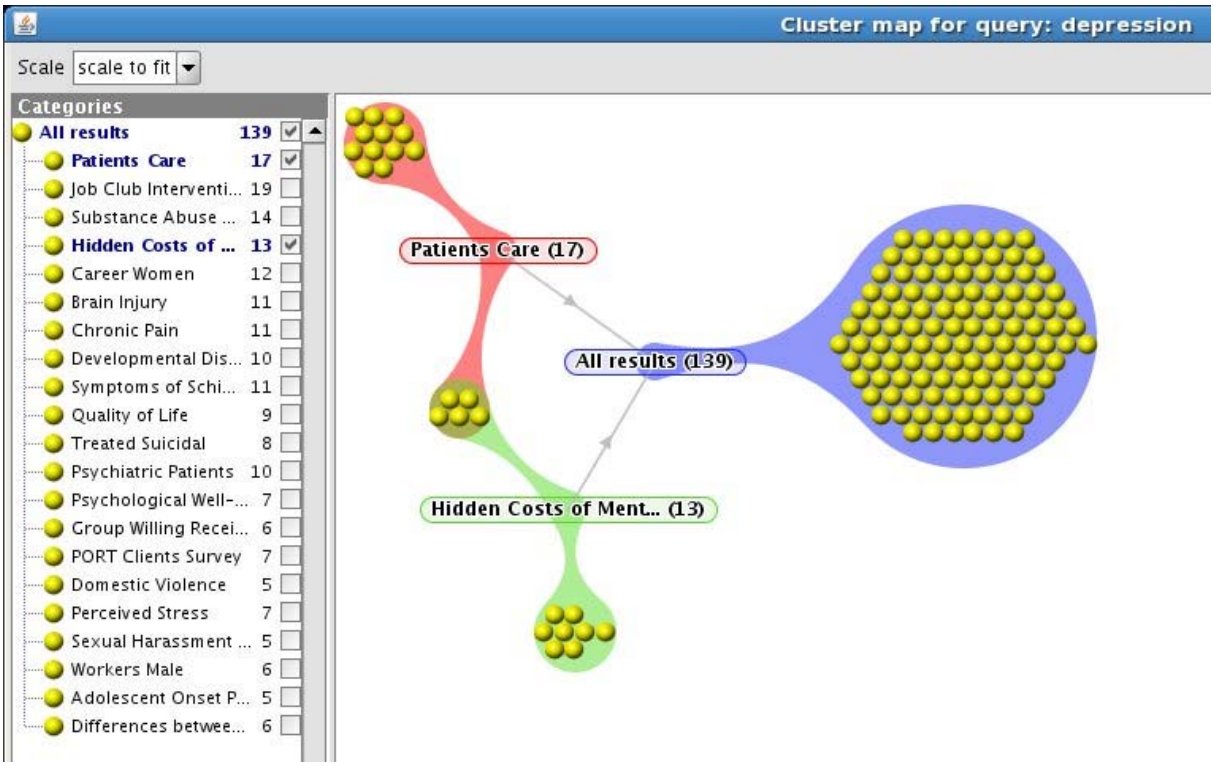

Figure 2: Visualisation of clustered documents for mental health systematic reviews

The current interface of ASSERT is based on the open source clustering engine Carrot2[10] (Osiński and Weiss, 2005b). Figure 2 shows an example of document clustering visualisation using Carrot2 for the domain of mental health. The clusters are 'Patients Care' and 'Hidden Costs of Mental Health' and the overlap between the sets are shown as a merged bubble. In this example the documents in the overlap are: *recovery from depression, work productivity and health care costs among primary care patients*; *cost effectiveness of practice-initiated quality improvement for depression*; and *gender patterns in cost effectiveness of quality improvement for depression.* By adding more topics, we gain a better overview of the documents in the collection. In addition, this offers the user a quick method of selecting only

[9] The most widely used measure is the tf-idf (term frequency–inverse document frequency) weighting scheme often used by search engines to score and rank a document's relevance given a user query.
[10] http://project.carrot2.org/demos.html

the documents that are of interest and can be used multiple times, by passing a cluster back into the system to see how that is re-categorised.

Clustering provides the main categories upon which we base the subsequent step, document classification. Where clustering groups documents together based upon similarity, classification examines the most important features that distinguish between the different clusters and use this to predict how a new document may be clustered in a much less computationally intensive manner.

Ongoing work includes query expansion based upon the extracted terms and their variants (e.g., acronyms, orthographic, morphological variants, and other synonyms).

Improving the screening strategy

One of the aims of screening is to narrow the scope of search, thus reducing the collection to only the relevant documents of a specific review. One approach we are using for this is topic classification to assist the user in focusing the review analysis on particular sub-topics within the overall review area. Document classification has been investigated by many researchers over the past 20 years (Osiński and Weiss, 2005b). In the late 1990s, machine learning techniques were successfully applied to topic classification (Dumais et al., 1998; Joachims, 1998) and after trials with other machine learning algorithms we settled on using support vector machines (SVMs) for its overall accuracy on the test domains.

As part of related research, we have been investigating the *Open Topic Assumption (OTA)* and *Closed Topic Assumption (CTA)* for document classification (Dumais et al., 1998). In most algorithms the OTA is used so that any topics not explicitly assigned to the document are treated neutrally, allowing for accurate results across many overlapping topics. The CTA, however, takes this further so that any topics not clearly predicted to a document *must* not be relevant and therefore creates new topic classes to account for the overlaps, providing more fine grained analysis. Through combinations of the two models it is possible to create systems that can classify the documents across a number of levels of detail or that can be customised to match the needs of specific domains based upon the scope, content and text type, such as questionnaires, interview transcriptions or academic reports.

Within this research we compared the document classification performance using a selection of potentially contributory features including:

- C-value terms: C-value terms that are extracted by using TerMine.
- tf*idf terms: uni-grams, bi-grams, tri-grams with tf*idf scores higher than a threshold.
- section: whether or not features are distinguished by document sections.
- negation: ignore bi-grams where the first word is negation words, such as *no*, *nothing*, or *not*.

As a preliminary experiment, we used the Medical NLP Challenge Data Set[11]. Our document classification system was trained on a data set with about 1,000 short anonymized clinical

[11] http://www.computationalmedicine.org/challenge/

records and their ICD-9 clinical codes.[12](Sasaki, Rea and Ananiadou, 2007). The system performances were evaluated on a further 1,000 unseen clinical records. The performance was measured by standard micro-average F1-measure and the multi-topic accuracy. In Table I, we see the results of our clinical document classification system based on the closed and open topic assumption approach. The experiments confirmed that the combination of C-value terms, tf*idf terms and negation archived the highest F1 score of 0.8672. The nature of the techniques used and features described for this experiment suggest that this is domain independent and could be adapted after initial training to other areas providing an efficient and effective means of assisting the challenge of screening across large document collections.

Approach		tf*idf + negation	tf*idf + negation + C-value terms	tf*idf + negation + section
Closed Topic Assumption	Micro-average F1	0.8322	0.8306	0.8417
	Multi-Topic Accuracy	0.7541	0.7520	0.7643
Open Topic Assumption	Micro-average F1	0.8634	0.8672	0.8594
	Multi-Topic Accuracy	0.7859	0.7900	0.7848

Table I: Evaluation results of document classification on clinical data

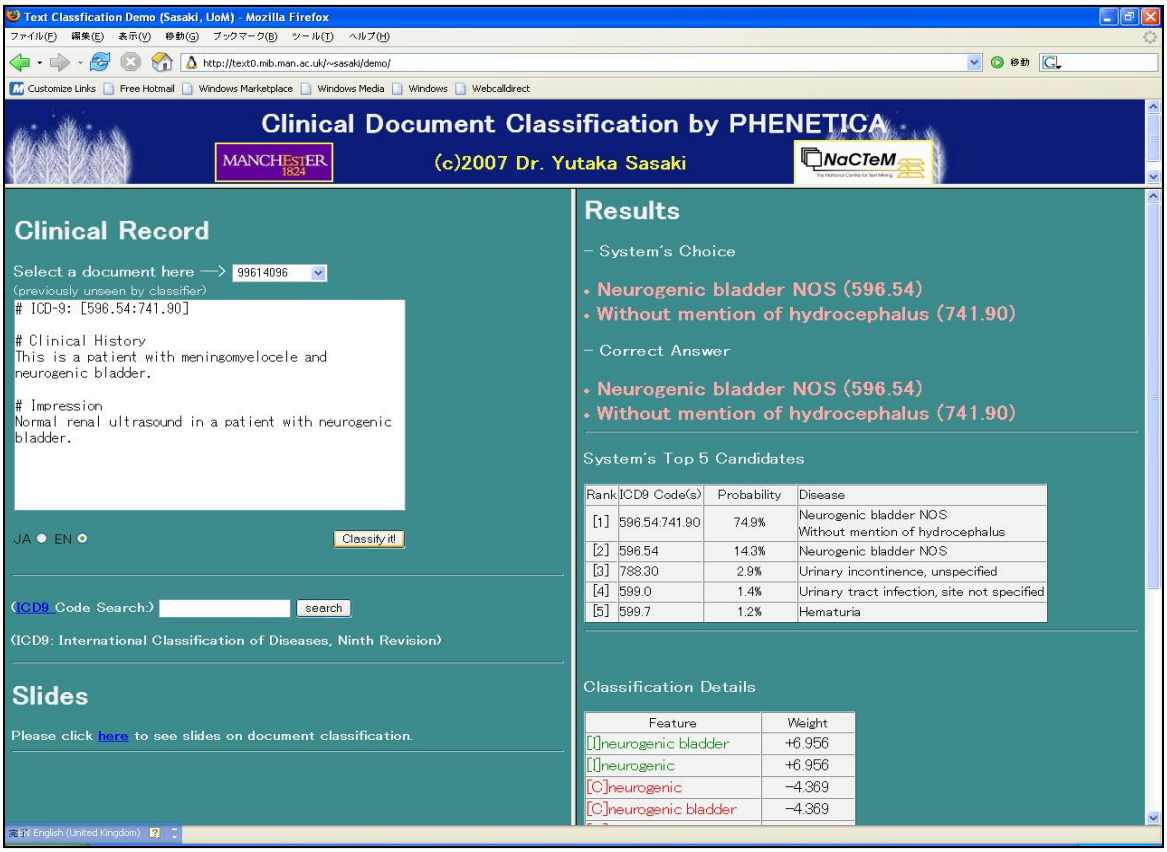

Figure 3: Screenshot of the document classification system

[12] http://www.cdc.gov/nchs/datawh/ftpserv/ftpicd9/ftpicd9.htm

Figure 3 is a screenshot of our document classification system[13] being applied to clinical records. When a record is processed the user is presented with the system's chosen topics, a correct answer based upon the known results (not presented to the classification system) and a list of the top five candidates with associated confidence. The combination of these results could provide an audit trail for the reviewer to appraise quantified evidence of the classification and the original document source, should a result ever be questioned. This is a key feature in technology assisted systematic reviews as trust in the results is vital for an effective and accurate review, also the direct linking between evidence and source can speed up any synthesis.

Aside the initial results of the classification system, it is possible to investigate the underlying features of what makes up a class. The features in these experiments were the set of terms, phrases and n-grams used as input to the classifier. In the system described above these features are presented in a ranked list of how they contribute to the overall classification result. By examining this data closely it is possible to gain an insight into how the terms and topics are related suggesting to the reviewer areas that may be appropriate for further investigation and also potentially identifying areas where further exploration may not be as fruitful.

Conclusions and further work

Using semi-automated techniques to perform some of the more time consuming tasks of systematic reviewing, reviews will be completed more quickly and importantly more systematically as more evidence from data will be harvested, filtered and summarised. In addition, searching, screening and synthesising will become more customised focusing on pertinent terms, retrieving relevant documents and synthesising salient information fragments. Critical aspects for the uptake of text mining technologies and tools in systematic reviewing are robust, scalable, efficient and rapidly responsive services for very large collections and the need to consult large-scale resources (corpora, thesauri). Equally important is the question of what is the right balance between automation of the process and user intervention and control. Our close collaboration with EPPI-Centre ensures that these issues will be thoroughly investigated.

In recent years, developments in e-Infrastructure have opened up new opportunities for the application of text mining applications and services (Carroll, Evans and Klein, 2005). Computationally intensive tools have previously only been usable on small scale tasks but are now being developed for much larger scale tasks thanks to alternative models of processing and storage. This allows us to expand on current tools to take into account the additional information available in full text documents and not just relatively small abstracts. With recent research showing that abstracts alone contain less than half of the overall information content of a paper (Corney et al., 2004), this is a significant boost for the analysis of documents. Combining large scale document repositories with web crawling technologies to provide access to the increasing amount of grey literature can offer vital insights into current research, potentially months before publication through traditional routes.

In all this provides growing opportunities for the application of text mining in systematic reviewing and in the social sciences in general. Text mining techniques have the potential to revolutionise the way we approach research synthesis, but our longer term interest is to understand how we can apply these techniques more widely in the social sciences. To achieve

[13] http://text0.mib.man.ac.uk/~sasaki/demo

this, we will use systematic reviewing to demonstrate the potential of text mining for the social science research community (Rea and Ananiadou, 2007) and to establish requirements for a generic toolkit of text mining services which can be integrated into different research practices.

This provides its own set of issues for development in terms of interoperability with techniques or software currently employed in the systematic review activity but also with other text mining tools and services used by the social science community. For example, a researcher investigating the role of new media in politics could be interested in combining the toolset with internet news feed or blog readers, their own evidence tracking systems or even other tools for carrying out opinion analysis. We need to ensure that our tools are therefore flexible and robust enough to allow for this, whilst providing sufficient functionality to ensure interoperability between the many formats and standards that this would entail.

Acknowledgements

The ASSERT project is funded by the UK Joint Information Systems Committee as part of its e-Infrastructure programme.

References

Anderson, R. (1994). Representations and requirements: The value of ethnography in systems design. Human-Computer Interaction 9, p. 151-182.

Bollegala, D., Okazaki, N. and Ishizuka, M. (2006). A Bottom-Up Approach to Sentence Ordering for Multi-Document Summarization, in Proceedings of ACL, pp.385-392.

Carroll, J., Evans, R. and Klein, E. (2005). Supporting Text Mining for e-Science: the challenges for Grid-enabled Natural Language Processing. UK e-Science All Hands Meeting, Nottingham, UK.

Corney, D.P.A., Buxton, B.F., Langdon, W.B. and Jones, D.T. (2004). BioRAT: extracting biological information from full-length papers, Journal of Bioinformatics, Oxford Journals, 20(17):3206-3213.

Chalmers, I. (2003). Trying to do more Good than Harm in Policy and Practice: The Role of Rigorous, Transparent, Up-to-Date Evaluations. *The ANNALS of the American Academy of Political and Social Science*, Vol. 589(1), 22-40.

Dumais, S., Platt, J., Heckerman, D. and Sahami, M. (1998). Inductive Learning Algorithms and Representations for Text Categorization, *Proc. CIKM '98*, pp.148-155.

Frantzi, K., Ananiadou, S. and Mima, H. (2000). Automatic Recognition of Multi-word Terms. *International Journal of Digital Libraries* 3(2), 117–132.

Hartswood, M., Procter, R., Rouncefield, M, Slack, R. and Voss, A. (2007). Co-realisation: Evolving IT Artefacts by Design. In Ackerman, M., Erickson, T. and Halverson, C. (Eds.) Evolving Information Artefacts, Springer.

Hartswood, M., Jirotka, M., Procter, R. et al. (2005). Working IT out in e-Science: Experiences of requirements capture in a HealthGrid project. In Proceedings of the HealthGrid Conference, Oxford.

Hey, A. and Trefethen, Anne. (2003). The Data Deluge: An e-Science Perspective. In Berman, F., Fox, G. and Hey, A. (Eds.) Grid Computing – Making the Global Infrastructure a Reality, Wiley.

Jirotka, M. and Goguen, J. (Eds.) (1994). Requirements engineering: Social and technical issues, London : Academic Press.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proc. of 10th European Conference on Machine Learning (ECML-98), pp.137–142.

Lin, C.Y. and Hovy, E. (2001). Neats: a multi-document summarizer. *Proceedings of the DocumentUnderstanding Workshop(DUC).*

Osiński, S. and Weiss, D. (2005a). A Concept-Driven Algorithm for Clustering Search Results. IEEE Intelligent Systems, May/June, 3 (vol. 20), 2005, pp. 48—54.

Osiński, S. and Weiss, D. (2005b). Carrot2: Design of a Flexible and Efficient Web Information Retrieval Framework. Springer Lecture Notes in Computer Science, vol. 3528, pp. 439—444, Proceedings of the third International Atlantic Web Intelligence Conference (AWIC 2005), Łodź, Poland.

Rea, B. and Ananiadou, S. (2007). Text Mining Services to Support e-Research. UK e-Science All Hands Meeting, Nottingham, UK, September.

Sasaki, Y. Rea, B. and S. Ananiadou, S. (2007). Multi-Topic Aspects in Clinical Text Classification. IEEE International Conference on BioInformation and BioMedicine, Silicon Valley, November.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorisation, ACM Computing Surveys, Vol. 34, No.1, pp. 1-47.