WORKING PAPER 03/01

Methodologies for the Automatic Location of Academic and Educational Texts on the Internet.

Linda Oxnard and Andrew Evans

School of Geography

University of Leeds

LS2 9JT

January 2003

ABSTRACT

Traditionally online databases of web resources have been compiled by a human editor, or though the submissions of authors or interested parties. Considerable resources are needed to maintain a constant level of input and relevance in the face of increasing material quantity and quality, and much of what is in databases is of an ephemeral nature. These pressures dictate that many databases stagnate after an initial period of enthusiastic data entry. The solution to this problem would seem to be the automatic harvesting of resources, however, this process necessitates the automatic classification of resources as 'appropriate' to a given database, a problem only solved by complex text content analysis.

This paper outlines the component methodologies necessary to construct such an automated harvesting system, including a number of novel approaches. In particular this paper looks at the specific problems of automatically identifying academic research work and Higher Education pedagogic materials. Where appropriate, experimental data is presented from searches in the field of Geography as well as the Earth and Environmental Sciences. In addition, appropriate software is reviewed where it exists, and future directions are outlined.

KEYWORDS

web, internet, search, automatic, content analysis, harvesting, academic

ACKNOWLEDGEMENTS

This report was funded by the Higher Education Funding Council for England Learning and Teaching Support Network Subject Centre for Geography, Geology and Earth Sciences. The paper builds, in part, on ongoing unfunded Ph.D. work at the University of Sheffield.

1. INTRODUCTION

The last ten years have seen considerable advances in the usability of systems for the production and distribution of hypertexts[1] with embedded multimedia components. Chief amongst these advances have been the developments associated with the World Wide Web (hereafter WWW or "web"). Most schoolchildren above Year 7 now have the skills necessary to write hypertext documents using the HyperText Markup Language (HTML) and publish these as 'webpages' on an Internet site. Most are also familiar with using computer-aided learning resources and utilising the Internet in research.

This development, in students at all levels, has been matched and led by a concomitant development of skills in the teaching and research communities. The ease of HTML use, and the production of authoring software, have led to a significant shift in the development of computer-aided teaching and research resources. A decade ago the development of almost all such resources required considerable programming ability. Now educators who need only know how to operate a basic word processor and graphics package in order to develop the same resources.

These changes are having a twofold positive effect on the teaching and learning process. Firstly, they are encouraging the development of students who are more critical - both of the information given and the learning process. Secondly, they are allowing lecturers in a given field to get a better overview of their teaching community at a global level and allowing them access to resources produced by others to aid teaching.

However, as the level of teaching and learning information, as well as research papers and project descriptions increases, there is a negative side effect, in that finding resources in any given subject area becomes harder. While experience-led improvements in choosing search-terms rapidly leads most users to find information in popular research fields using search engines such as Google[2], there are a number of factors mitigating against other academic materials coming out at the top of any given search results.

[1] That is, texts in which words or phrases are linked to additional or related information.
[2] http://www.google.com

1) Many scientific communities are small. For sites like Google, which rate pages on the amounts of links to them[3], this can have a negative impact. For sites that rate pages on the basis of their popularity, this can be even more devastating.

2) Lecturing communities are even smaller. It's unlikely that one lecturer will link to another's materials or visit them more than once. This is particularly true while the model of a course embedded in a (geographically fixed) degree scheme, written or managed by a single academic, is the dominant model.

The problems are particularly noticeable where the subject area covers information given a more popular treatment by non-academic groups, for example, the oil industry, conservation, materials on specific locations, economics or politics. In these cases, popular news and advocacy sites will dominate search results.

Because of this, the last five years has seen the growth of so-called 'portal' sites that provide information on specific subject areas. The usual format of such sites is to tempt users with a number of online services and information sources, while supplying a database of links and resources as their chief utility contributed to the users. Academic examples include the Resource Discovery Network[4] and the Australian Subject Gateways Forum[5]

Traditionally such online databases of web resources have been compiled by a human editor, or through the submissions of authors or interested parties. The considerable resources needed to maintain a constant level of input and relevance in a world of increasing material quantity and quality, along with the ephemeral nature of much of the content of the web, dictates that many sites stagnate after an initial period of enthusiastic data entry. The solution to this problem would seem to be the automatic harvesting of resources, however, this would necessitate the automatic classification of resources as 'appropriate' to a given database through the difficult process of analysing the texts' content.

[3] For more information on Google's PageRank system, see http://www.google.com/technology/index.html
[4] http://www.rdn.ac.uk/
[5] http://www.nla.gov.au/initiatives/sg/gateways.html

This paper will outline the component methodologies necessary to construct such an automated harvesting system, including a number of novel approaches. In particular this paper looks at the specific problems of automatically identifying academic research work and Higher Education pedagogic materials. Where appropriate, experimental data is presented from searches in the field of Geography as well as the Earth and Environmental Sciences.

There are three key stages to creating an automated web portal capable of finding, classifying and categorising educational and academic material.

1) Text location: finding resources on the web for potential inclusion.

2) Style analysis: examining each potential resource to see if its origin is “academic”.

3) Subject analysis and classification: examining each potential resource for actual content followed by its placement in some easily navigated classification structure.

The first stage proceeds through the use of a focused crawl of the web looking for texts covering a particular subject area.

The second stage involves filtering with the aid of a stylistic identifier.

The third stage involves Part of Speech (PoS) tagging[6] of each text and an analysis on the resources to confirm their nature.

Plainly, if the first stage uses a subject classification there will need to be some interaction with the third stage as to how the system defines a reasonable classification. In addition, there are a number of ways in which the stages can be conflated to increase computational and searching efficiency. Given these interactions, this paper will also examine the ordering and linking of appropriate methodologies. In addition, useful software will be reviewed where it exists and, at the end of the paper, future directions in this area will be outlined.

[6]Part of speech tagging assigns text labels to all words within a document reflecting their syntactic category. For example, the sentence 'John kicked the ball angrily' would be tagged John (NP) kicked (VPT) the (ART) ball (NC) angrily (ADV) where NP = proper noun, VPT = past tense verb, ART = article, NC=common noun and ADV = adverb.

2. FINDING MATERIALS

2.1 Focused web crawling

"There is much awareness that for serious web users, focused portals are more useful than generic portals: the most interesting trend is the growing sense of natural limits, a recognition that covering a single galaxy can be more practical and useful than trying to cover the entire universe".

(Chakrabarti *et al.*, 1999)

There are at least 2000 million pages on the web[7]. Any portal seeking to catalogue, for example, educational geography texts, will find that they constitute a very small subset of the whole. It would be extremely wasteful in terms of resources to 'crawl' (scan through), 2000 million documents in order to find an extremely small fraction of them.

In conventional web crawling, the crawler software (or 'robot') is given a starting page, which it examines for its purposes before following all of the links from that page to subsequent pages. Each of these new pages is then examined and scanned for more links, which are also all followed for new pages, and even more links. In this way, a conventional web crawler can quickly find its way to thousands of pages, all of which are a within a certain number of 'clicks' (hypertext jumps) from the initial starting point. Any of those thousands of pages that are considered relevant during examination are recorded.

Focused crawling begins in the same way as a conventional crawl, by following all of the links from a specified starting page. However, as each subsequent page is retrieved, it is tested to see if it is a relevant resource. In a focused crawl, only the pages that are relevant are scanned for links to other pages for retrieval. (In some cases this rule is relaxed to allow crawling through a few non-relevant pages before stopping).

A detailed study of the performance of focused crawling was undertaken by Chakrabarti *et al.* (1999) who studied both focused and unfocused crawling on

[7] There are no reliable estimates of the current size of the web, however, there were 2,073,418,204 catalogued on the 6th July 2002 at www.google.com, and there is still considerable difference in the pages catalogued by this search engine when compared with others (Notess, 2002), suggesting this is a significant underestimate of the total pages in existence.

specific topics, with both searches starting from the same initial page. Figure 1 shows the results for an unfocused search: the vertical axis shows the average relevance of the pages retrieved whilst the horizontal axis shows the number of pages that have been examined. It is clear from the graph that, by the time only a thousand pages have been examined, the average relevance of those pages is almost zero.

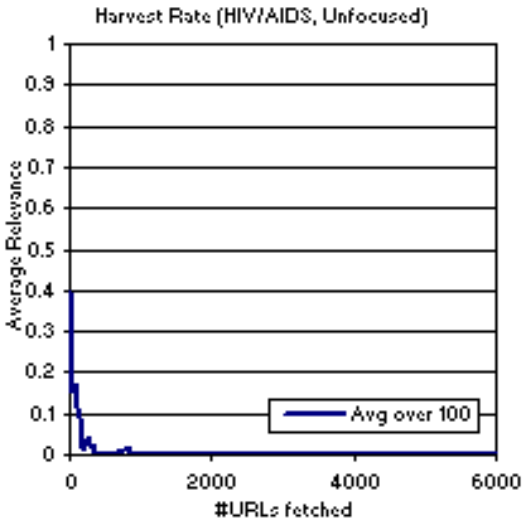

Figure 1: From Chakrabarti et al. (1999) - Rate of relevant page acquisition with a standard unfocused crawl on topic of HIV/AIDS.

By contrast, Figure 2 shows the results of a soft focused crawl (one which is allowed to follow links from less relevant pages a limited number of times). Here, rather than rapidly falling to zero, the level of relevance fluctuates but stays high, even after 5000 pages.

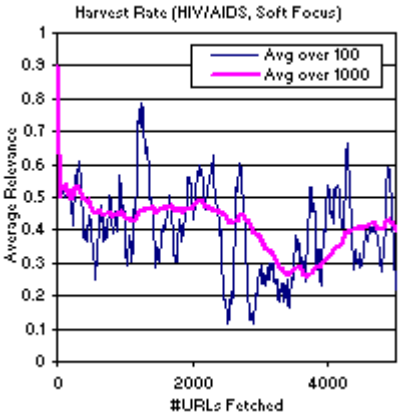

Figure 2: From Chakrabarti et al. (1999) - Rate of relevant page acquisition with a soft focused crawl on topic of HIV/AIDS.

Figure 3 shows the results of a hard focussed crawl (one which is only allowed to follow links from pages that are relevant) which still finds relevant pages after 10000 pages.

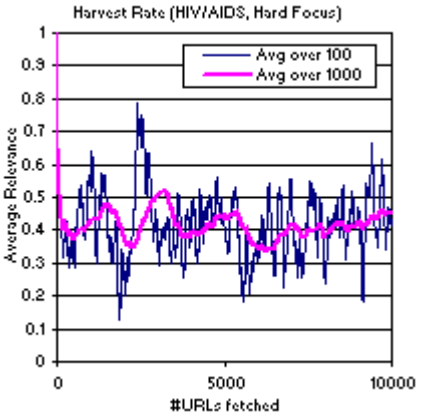

Figure 3: From Chakrabarti et al. (1999) - Rate of relevant page acquisition with a hard focused crawl on topic of HIV/AIDS.

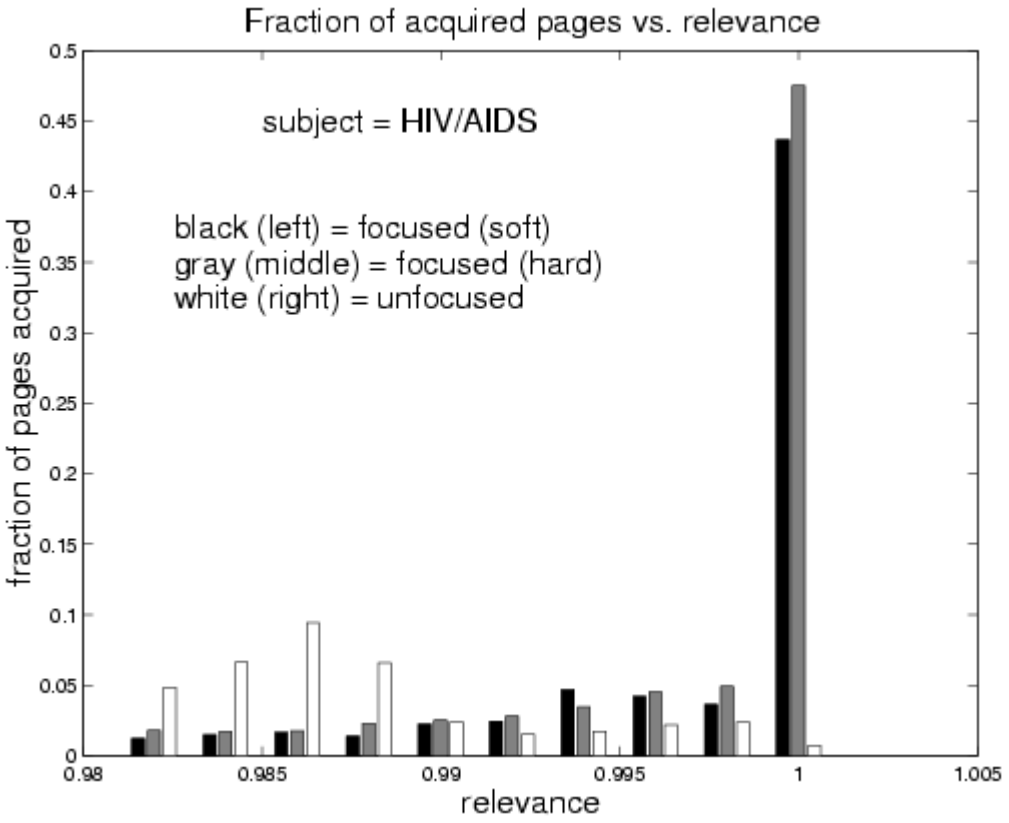

Figure 4: From Chakrabarti et al. (1999) - Distribution of relevance scores from the three crawlers.

Chakrabarti *et al.* show that pages obtained by focused crawling show a very sharp peak at the highest possible relevance value, whereas the unfocused crawler shows a fairly flat distribution of relevance (Figure 4).

In experiments, detailed below, between a third and half of all page fetches result in success for hard and soft focused crawlers.

Two types of hypertext mining programs usually guide crawlers: *classifiers*, that evaluate the relevance of a hypertext document with respect to the focus topics and *distillers*, that identify hypertext nodes or 'hubs' that are good access points to many relevant pages within a few links.

In Chakrabarti *et al.*, focused crawling acquires relevant pages steadily while standard crawling quickly loses its way, even though they are started from the same set of root pages. Focused crawling is capable of exploring out and discovering valuable

resources that are dozens of links away from the start set, while carefully pruning the millions of pages that may lie within this same radius. Focused crawling is very effective for building high-quality collections of web documents on specific topics, using modest desktop hardware.

The focused crawler achieves respectable coverage at a rapid rate because there is relatively little to do. Thus, in addition to finding resources, web content databases can also be maintained against depreciation by a distributed team of focused crawlers, each specialising in one or a few topics. Each focused crawler will be far more nimble in detecting changes to pages and assessing their continued relevance within its focus than a crawler that is crawling the entire web.

2.2 Tests

WebSPHINX[8] is a 'personal, customizable web crawler' created by Carnegie Mellon University that provides Java class libraries and an interactive development environment for web crawlers. Classifiers can be plugged into WebSPHINX to limit and direct searches.

Figure 5 shows a sample output from an unfocussed WebSPHINX crawl, in this case an unfocused crawl starting from an educational page about glaciers. From that point outwards, 100 sites were visited, only four of which were considered ‘on topic’ (using a crude test which simply looked for the occurrence of the word ‘glacier’ in each page’s text).

The starting page did contain links to other educational resources, but it also contained links to a dozen search engine home pages that caused the crawler to become lost in a large number of unrelated pages.

[8] http://www-2.cs.cmu.edu/~rcm/websphinx/

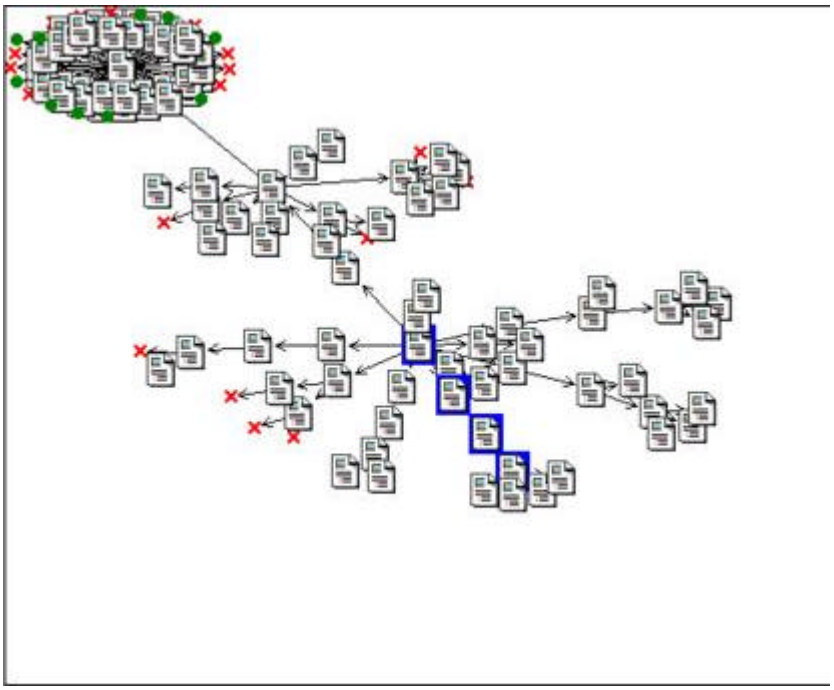

Figure 5: visual output from a WebSPHINX, unfocused crawl.Relevant pages are in blue.

Figure 6, on the other hand, shows another output from a WebSPHINX crawl: this time using a very crude focusing approach. Only links from pages containing the word 'glacier' were expanded. In a production system, a far more accurate classifier, similar to the ones discussed later in this report, would be used to identify and rank potentially ‘on topic’ pages.

Even with this crude focusing, the number of ‘relevant’ sites within the first 100 pages has jumped from 4 to 26.

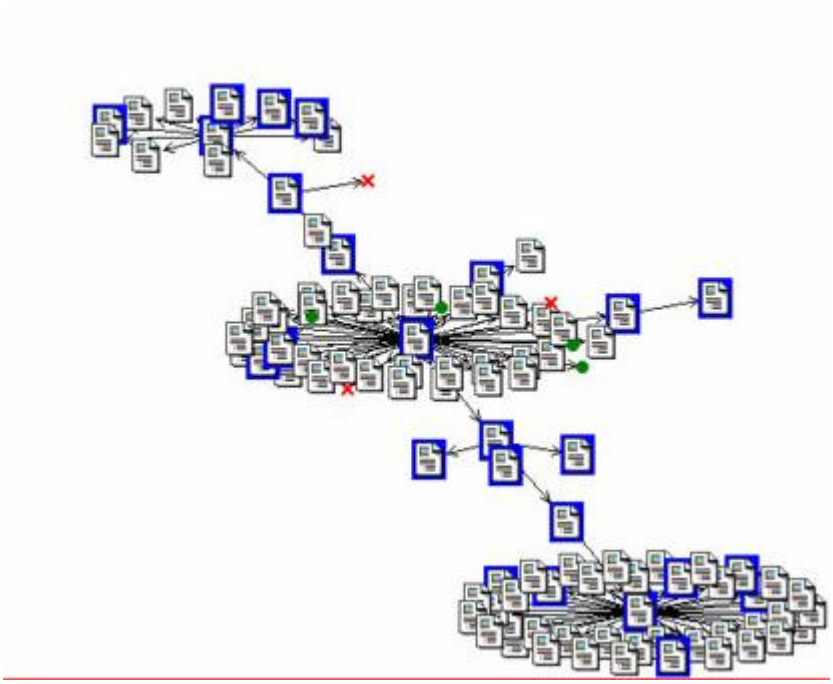

Figure 6: visual output from a focused, WebSPHINX crawl. Relevant pages are in blue.

2.3. Establishing a starting point

A key decision when performing crawls is the selection of starting pages. The three main possibilities are 1) a large search engine such as Google (feed search queries into Google to find pages and then start the crawl from the results page), 2) Hub pages, or pages which contain a number of academic related links such as an existing academic portal or gateway, 3) 'links' pages from relevant university departments (a soft focused crawl going to a depth of 5 or 6 links). If crawls are started on highly relevant pages they are perhaps less likely to pick up rogue texts. These are pages which contain identified keywords but in entirely the wrong context. As an example, consider the online paper "Full-band-structure theory of high-field transport and impact ionization of electrons and holes in Ge, Si, and GaAs" (Fischetti *et al.*, 1996) which contains both the word "avalanche" and the word "shear": keywords for "avalanche" texts in the field of glaciology (see below, Section 4.6.2). A conventional search engine is unable to distinguish between the use of keywords in the field of semi-conductor physics against those in a geographical context. Focused crawling avoids this problem, as pages connected to relevant pages are unlikely to be in a radically different field whilst using the same keywords.

3. STYLISTIC ANALYSIS

3.1. Register Analysis

There are many ways of saying the same thing in a given language; for example, one might speak 'formally' or 'casually' depending on the situation. Such "ways of speaking" are known as language "registers"[9]. In identifying academic texts we are looking for formal information of an academic content.

The remit of this paper is learning resources, whether this is in the form of educational material or academic texts. In principle, this means that it does not consider business pages, marketing pages, personal home pages, fan pages or a host of other styles of page, each with a different register from academic materials.

Register analysis is necessary in order to determine whether a text fits the description of 'academic paper' or general 'educational material'. Academic *research papers* are actually relatively easy to identify, as they tend to follow a fairly rigid set pattern in terms of their headers. These traditionally consist of two or more of the following: abstract, keywords, introduction, results, conclusions, references.

The ability to automatically identify more general *educational resources* is more complex. Firstly, before an attempt can be made to create rules for automatically identifying this material it is necessary to determine exactly what this material should consist of. For example, many 'academic' portals claim to include educational material, but much of this is made up of power point slides or notes that accompany real world lectures. Whether this sort of material is intended to be included or whether more emphasis is required on teacher notes or student aids needs to be clearly understood before register analysis can take place as, in most text analysis situations, a body of sample material (a 'corpora') is required to be gathered before any analysis can take place or conclusions formed.

[9] See, for example, ISO (1999), for a list of registers one is likely to encounter in the computer analysis of texts.

Kessler *et al.* (1997) have suggested four groups of generic cues that help in identifying text genre:

1) Structural cues: examples of structural cues are passive constructions[10], nominalizations[11] and syntactic category markers (part of speech tags).

2) Lexical cues: examples of lexical cues are Latinate affixes which signal certain highbrow registers or words used in expressing dates, which are common in certain types of narratives such as news stories.

3) Character-level cues: examples of character-level cues are punctuation marks and other separators used to mark text categories such as phrases, clauses and sentences, in addition to capitalised words and acronyms.

4) Derivative cues: examples of derivative cues are ratios and variation measures derived from measures of the features from the above three categories, for example, average sentence length, average word length, token/type ratio[12].

Kessler *et al.* identify two key areas of register analysis: *Brow* and *genre*. *Brow* characterises a text in terms of the presumptions made with respect to the required intellectual background of its target audience and is measured as *popular, middle, upper-middle* and *high*. For example, a copy of The Sun newspaper might be described as *popular*, The Guardian as *middle*, The Financial Times as *upper-middle* and an academic research paper as *high.*

Genre characterises a text in terms of its content. Examples of *genre* as defined by Kessler are *reportage, editorial, scitech, legal, nonfiction* and *fiction.* As *genre* analysis has the potential to automatically differentiate texts in terms of their contents, it is of particular use at the stage of distinguishing educational material from non-

[10] Passive voice is a voice that indicates that the subject is the recipient of the action denoted by the verb. For example, “The cat was seen by the dog” is the passive form of “the dog saw the cat”.

[11] The creation of a noun from a verb or adjective. A strong feature of written texts, nominalizations typically end in -ity, -tion, or -ness e.g. kindness (from the adjective 'kind'), density (from the adjective 'dense'), negation (from the verb 'negate'), etc...

[12] Token/type ratio is the ratio between the total number of words in a text and the occurrences of different words. For example, the sentence "I gave my friend a present, but my friend did not like the present and gave it back to me". There are 20 words (tokens) in this sentence, but only 16 different words, so the token/type ratio is 1.25. The closer the token/type ratio is to 1, the more complex the text is in terms of different words used.

educational material. In addition, *brow* analysis provides the ability to further classify the educational material by identifying the intended audience of the material and thereby the level of education to which it is addressed (i.e. pre-school, primary school, senior school or university).

Kessler *et al.* (1997) have performed separate experiments to analyse genre and brow based both on surface cues (i.e. derivative, character-level and lexical cues) and structural cues. It is interesting to note that they obtained largely comparable results for both methods. Indeed, they argue that there is at best a marginal advantage to using structural cues in brow and genre analysis work, an advantage that, in most cases, would not justify the additional computational cost required.

Levels	**Surface cues**	**Structural cues**
Genre		
Reportage	75	79
Editorial	96	93
Legal	96	93
Scitech	100	93
Nonfict	67	73
Fiction	93	96
Brow		
Popular	74	72
Middle	66	58
Uppermiddle	74	79
High	84	85

Table 1: table from Kessler et al. (1997) showing percentage of texts correctly identified according to brow and genre using structural and surface cues.

Biber *et al.* (1998) claim that groups of *co-occurring* features are instrumental in distinguishing among registers, that is, such texts may be identifiable by the statistical distribution of pairs of words, word fragments, or characters (letters, punctuation). For example, educational material might have more second person pronouns (you / your) in conjunction with commands (explain, write, read, compare) along with question words (what, why, where). Biber *et al.*'s recommended methodology is to apply such multi-dimensional analysis, and identifying characteristic co-occurrence patterns quantitatively, using a corpus of texts. They advise this approach as they claim that there is "no way of knowing ahead of time which individual features will be important in any given register analysis".

The creation of rules for distinguishing registers in educational or academic material is best be achieved through analysis of PoS tagged documents, however, this approach is extremely prohibitive because of the large amounts of computing processing necessary. In addition, this would require considerable human intervention in order to sort out useful from non-useful documents. Therefore, a more appropriate methodology is to look at the 'surface' features of a text, such as exclamation marks, question marks, capitalisation, sentence length, word length etc.

Counts of potential distinguishing factors are created and statistical analysis is applied to identify linguistic feature co-occurrence. For example, in work done on Catalan texts by one of the authors (Oxnard, in prep.), counts were created for punctuation marks, nouns, verbs, prepositions, personal pronouns, different verb tenses, etc. and these were compiled into two functions which gave a clear classification of three text types (Figure 7). Given this success it is felt that it could potentially be of some use in classifying academic and/or educational material.

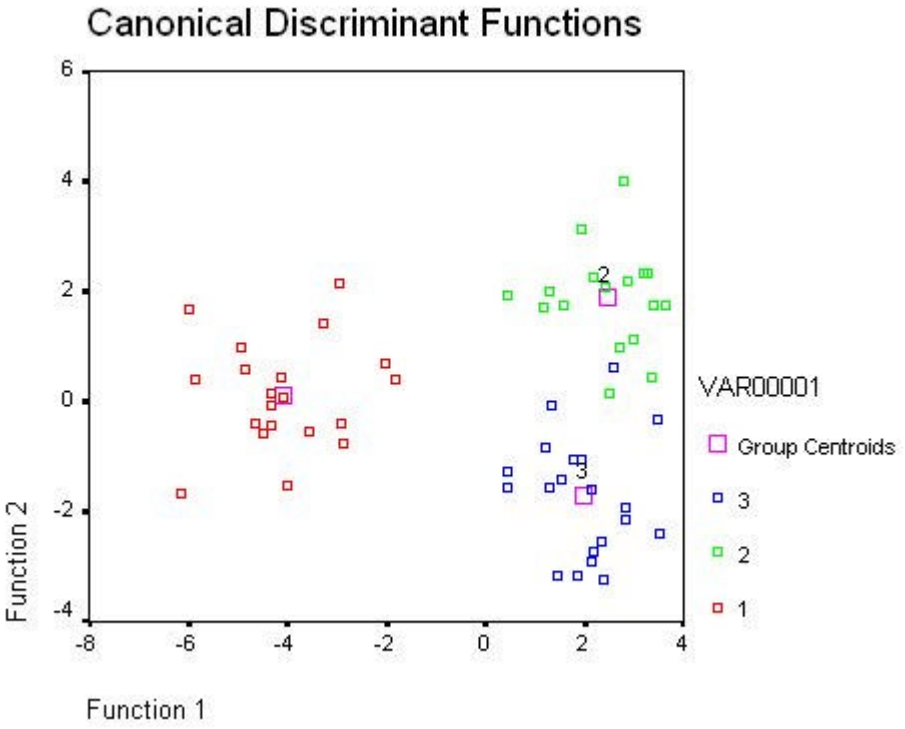

Figure 7: text categorisation using multi-dimensional analysis techniques.

3.2 Text quality

It is perhaps worth pointing out at this stage that whilst register analysis makes possible the identification of different styles of writing such as academic prose or newspaper reportage, it does not make possible an assessment of the actual quality of individual texts. This is a very subjective judgement: one that would not necessarily achieve agreement by more than one human judge. Beyond the recognition of authors who contradict themselves or others, there is little a machine can currently contribute to this process (see, for example, Iwanska and Shapiro, 2000). Distinguishing 'good quality' material from 'bad quality' material is a task best left to humans (although it must be recognised that humans themselves are unlikely to show high rates of agreement in performing such a task).

The role of the machine is to help find relevant material for subsequent review by a human editor / editors. That said, automated systems of peer review are possible, either by allowing end users of a system to rank material that they read, or through citation/reference counting, in which pages that are linked to by a large number of other pages are regarded as likely to be of more interest and/or of higher quality then those which are linked to infrequently. The search engine Google uses this principle

in its 'Page Rank' technology.

4. SUBJECT ANALYSIS

Ideally, any system identifying academic or educational texts will also classify the subject area the documents fall into. This can clearly be viewed as a text classification task and it is the topic of subject analysis that will form the main focus of this paper.

There are two main fields that are particularly relevant to automated text categorisation and classification tasks: Information Extraction (IE) and Keyphrase Extraction (KE).

4.1 Information extraction

Information extraction involves extracting specific types of task-dependent information from a document[13]. Whilst IE technologies automatically extract very detailed information, they are rule-based and require a large number of work hours for experts to set them up. In addition, owing to their strict rule bases, they cannot subsequently be ported to another domain.

Their main advantage over other types of system is that they achieve high precision because their approach is context sensitive. By examining context, IE systems are able to classify texts that would be impossible to classify using other techniques because they do not contain any keywords or keyphrases (not that this is likely to be a problem in academic / educational texts).

The major disadvantage to IE systems is that they work using a knowledge-based approach relying on a domain-specific dictionary. Such full-blown Natural Language Processing (NLP) systems are generally very expensive and can seriously strain computational resources.

Furthermore, it is felt that it would be wiser to use approaches that do not rely on pre-existing dictionaries and word banks for two reasons. Firstly, it means the system is

[13] One particular example that is often quoted in IE papers is that of extracting detailed information from news reports relating to terrorist activity. Using IE techniques it is possible to automatically locate detailed information from such articles, including the name of the terrorist organisations that carried out attacks, the names of the victims, the type of weapons employed etc. For the uses of such mechanisms, see, for example, Hunt (1996).

not limited to one national language. Secondly, academic fields of knowledge are in a constant state of flux. New knowledge and techniques are discovered all the time. If a system is trained on existing word banks and dictionaries, it becomes unable to spot new and relevant fields of interest.

Finally, IE is not suitable for the task in hand both because of the long set up period it requires and also because the level of detail it can provide is not required for general text classification work. In summary, the technique does not justify the computational resources it requires.

4.2 Keyphrase / keyword extraction

> *Keyphrases provide a powerful means for sifting through large numbers of documents by focusing on those that are likely to be relevant*
>
> Frank *et al.* (1999)

Keyphrase extraction examines a text and automatically extracts those words contained within the text that it considers to be the most important. Turney (2000) describes automatic keyphrase extraction as "the automatic selection of important, topical phrases from within the body of a document". Keyphrase extraction is not as specific as IE, but as it is fully automated and non-rule-based, it does not require a huge amount of expert labour to make it work.

The importance of high quality keyphrase extraction for fuelling an automated text classification system can be seen by observing how humans classify documents. Humans can quickly and easily pick out relevant documents by skim reading them and pulling out relevant words, hence making a preliminary survey of the text and its contents. When skim reading a text, a human is capable of quickly locating 'keywords' or 'keyphrases' within that article which hold vital clues as to the field from which it came. For example, a section from an academic article in the field of glaciology has been replicated below (Hodson, 1999). Keywords and phrases that tie the text to its field have been underlined.

Investigations from Svalbard over the last 10 years have contributed significantly to the number of glacio-fluvial process studies conducted in high Arctic basins (e.g. Barsch *et al.*, 1994; Bogen, 1991; Hodgkins, 1996; Hodson *et al.*, 1997; Hodson *et al.*, 1998; Kostrzewski *et al.*, 1989; Repp, 1988;

Vatne *et al.*, 1992). However, despite this advance, our understanding of the linkage between glacier hydrology and proglacial sediment and solute transfer remains dominated by research conducted within temperate glacier basins. Most of the temperate glacier research has argued that temporal changes in proglacial discharge, sediment and solute fluxes are caused by changes in the mixing ratio of two or more reservoirs with contrasting residence times, pathways and degrees of rock:water contact (Collins, 1977; Oerter *et al.*, 1980; Gurnell and Fenn, 1984a; Sharp, 1991; Fountain, 1992; Tranter *et al.*, 1993, 1997; Clifford *et al.*, 1995a; Richards *et al.*, 1996). Of critical importance for the timing and magnitude of these changes is the combined evolution of two subglacial reservoirs: an efficient channelised reservoir with short residence times and supplied predominantly by icemelt, and an inefficient, highly distributed reservoir, supplied predominantly by snowmelt (Richards *et al.*, 1996; Tranter *et al.*, 1996; Willis *et al.*, 1996). Typically, the co-evolution of these two reservoirs throughout the ablation season is believed to involve an increase in the extent of the glacier bed drained by the channelised system at the expense of the distributed system (e.g. Richards *et al.*, 1996, Iken and Truffer, 1997).

As well as being those words which 'jump out of a page' and inform a reader as to the field of knowledge to which a document belongs, they are also the words which are capable of informing an automated system as to the subject of a text.

Keyphrase extraction would appear to be the key to automated text classification as it greatly simplifies the task of content identification and classification. Rather than determining the subject of an article from the entire text, a classification system would only have to work with the key words identified by the extractor as being the most important.

For the purposes of classification, classes that match keyword lists may be enough. Alternatively, keywords could be used as the input to an inductive machine learning technique that would generate topic areas from keyword sets.

4.3 'Recall' and 'precision'

Two key concepts to the field of keyphrase extraction are *recall* and *precision*. *Recall* measures the percentage of relevant texts that are correctly classified as relevant. *Precision*, on the other hand, measures the percentage of classified texts that are correctly relevant.

To illustrate these concepts, imagine a hypothetical set of 200 texts, of which 100 are geography related. A KE system is set up to locate and categorise geography texts

within this set. A system with a high recall rate might return 120 geography texts (90 true geography texts plus 30 which are, in fact, non geography texts). A system with a high precision rate, on the other hand, might return 70 texts (all of which would be true geography texts). High recall means locating as many appropriate texts as possible, high precision means making sure that the texts located are all genuine. As can be seen, the best system is one that combines a high recall rate with a high precision rate. However, recall and precision levels are generally inversely proportional. Consistent high precision is often only possible at relatively low recall levels.

4.4. Author keywords

One question we might ask is why we should go to the trouble of extracting keywords from texts when, particularly in the case of academic research papers, these often form part of the text itself, in the shape of author assigned keywords. There are several reasons why author keywords do not prove particularly useful for locating further papers and articles on similar topics. Three of the most obvious reasons are discussed below.

Firstly, author keywords are sometimes added at the end of the writing process, with little thought, simply to comply with journal standards.

Secondly, authors often assign keywords to their papers not in order to make their work easier to find by interested parties, but rather to make their paper stand out in search engines or to show that their paper is relevant to the particular journal or publication to which they are submitting (even if, in fact, it is not particularly relevant: in which case the keywords often give very little indication as to the real thrust of the paper).

Finally, author keywords, as might be expected, do not generally incorporate words that might be considered to best describe the content of an academic paper. For example, an academic research paper on ‘Tyrolian avalanches’ is unlikely to include ‘snow’ as an author's keyword, even though the appearance of this word in the body of a text is one very obvious signal when looking for avalanche related papers.

In particular, with the final point made above in mind, even if a document does contain its own keywords that have been assigned by the author, it is still felt to be necessary to augment these keywords with other significant phrases that are included in the body of the text in order to create a body of keywords.

Once we have a body of keywords for texts we know are of interest, we can use these to locate other texts in the same subject areas.

4.5 Automatic keyphrase extraction techniques

A number of techniques exist for the automatic extraction of keyphrases from text. The following sections summarise how these techniques work and also introduce two working systems that are readily available.

4.5.1 Noun phrase[14] (NP) skimming

This method, outlined by Barker and Cornacchia (2000), involves choosing noun phrases based on their length, frequency, and frequency of their head nouns. To achieve this, it is first necessary to Part of Speech (PoS) tag each document in order to identify noun phrases. Once a noun phrase has been identified, the noun and adjective status of its words are checked in a dictionary. Keyphrases are then extracted using a NP skimmer and an online dictionary.

In particular, the length of NPs are taken into consideration because it is suggested that longer NPs with more premodifiers are more specific and may be more relevant to a particular document than more general, shorter, NPs. This is true, but unfortunately not of use here, because of the need to locate key words and phrases that are likely to appear in many other documents on a similar subject so we can use them in focussed searches. Tests counting word, bigram (two-word fragments) and trigram frequencies has shown that keyphrases made up of more than two words tend to produce fewer similar documents when searches are performed (Oxnard, in prep.).

4.5.2 Word positions and frequencies

[14] A noun phrase is a phrase that has a noun as its *head*, that is, that the noun is the single obligatory element in the construction. For example, in the phrase "wet paint", "paint" is the head noun. You can remove "wet" and the thing the sentence points out still makes sense, which is not true if just "wet" is left.

Some systems look for most frequently appearing words. This has disadvantages as, particularly in academic texts, synonyms are frequently used to avoid repetition, making simple word frequency counts of limited use.

Most automated keyphrase extraction systems use more complex algorithms to extract keyphrases. The software package *Extractor*, for example, scores candidate phrases on frequency of words with common roots in the phrase, length of phrase, and position of phrase in document.

4.5.3 Structural features

Krulwich and Burkey (1996) extract keyphrases from documents based on the structural and superficial features of the document. They use several heuristics including focusing on phrases which appear in section headers and phrases which are formatted differently from the surrounding text. They claim that words that appear in italics or bold are often words that are important to the text. However, this is not always the case, as words often appear in different typefaces for a number of other reasons such as emphasis, or signalling non-native or unfamiliar words.

4.5.4 Synonym dictionaries

Repetition is a major clue for KE systems that a candidate phrase is a keyphrase. Due to the frequent use of synonyms to avoid repetition in academic text, some system creators believe that results of automated keyphrase extraction could be significantly improved by adding synonym detection to the keyphrase extraction algorithm (some researchers have already attempted this using WordNet[15]). However, the use of synonym dictionaries makes the process language dependent, more computationally expensive and less robust.

4.6. A brief overview of two existing systems

This section considers two existing automated keyphrase extraction systems, their efficiency, methodology, required computing resources, training periods, and how

[15] WordNet is an online lexical reference system. Nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. WordNet can be found at http://www.cogsci.princeton.edu/~wn/.

difficult it would be to apply them to the task in hand.

4.6.1 Extractor 7.0: http://extractor.iit.nrc.ca/on_line_demo.html

Extractor is a keyphrase extraction system created Peter Turney at the National Research Council of Canada. It uses a genetic algorithm to calculate keyphrases and is capable of extracting between 3 and 30 keyphrases from any document with which it is presented in English, French, German, Spanish, Japanese or Korean. It is possible to extract more than 30 keyphrases if necessary by passing the required text into Extractor in smaller chunks. The software also offers the ability to combine keyphrases that have been extracted from a number of documents.

While the complete source code to Extractor is available from the National Research Council of Canada it is not free. The software is constructed as a Dynamically Linked Library (DLL) with an Application Programming Interface ('API') so that it can be easily embedded into other software. It is also capable of directly handling HTML files.

4.6.2 Kea: http://www.nzdl.org/Kea

Kea is a keyphrase extraction system, available under the GNU public license, and developed by The New Zealand Digital Library Project, a research programme at the University of Waikato (Witten, *et al.* 1999). The software builds on Turney's work with Extractor but extracts keyphrases using a Bayesian approach instead of a genetic algorithm approach.

The way in which Kea selects keyphrases is by using an algorithm incorporating the position of a word's first occurrence, and how often a word appears in a particular document (the Term Frequency) compared against with how often it appears in a global corpus (the Inverse Document Frequency).

The authors suggest that Kea's performance can be boosted significantly if it is trained on documents that are from the same domain as those from which keyphrases are to be extracted. They claim this allows the user speedier training than Extractor and that deriving such domain-specific models is less practical with genetic algorithm approaches. In fact, experiments run using Kea on academic materials seem to suggest

that training on relevant documents achieves only marginal benefits as can be seen in Tables 2 and 3 below. Whether the marginal improvements in keyphrase quality are worth the effort required to locate and train appropriate text types seems questionable.

Kea's performance is said to be close to optimum if about 50 training documents are used. 50 documents were used in these experiments.

Manually picked keywords	**No. Of Kea matches**	**Author keywords**	**No. Of Kea matches**	**Computer Science keywords**	**No. Of Kea matches**
aquifers	0.6	aquifers	0.6	classification	0.7
classification system	0.6	classification system	0.6	aquifers	0.5
groundwater	0.6	groundwater	0.3	groundwater	0.5
aquifer classification	0.6	aquifer classification	0.3	classification system	0.5
groundwater management	0.6	aquifer classification system	0.3	aquifer classification	0.2
aquifer classification system	0.6	ranking values	0.3	aquifer classification system	0.2
System for Groundwater	0.4	vulnerability	0.3	vulnerability	0.1
vulnerability	0.4	aquifer classes	0.3	ranking values	0.1
ranking values	0.4	British Columbia	0.3	water	0.1
British Columbia	0.4	Fraser River	0.1	map	0.1

Table 2: Comparing extracted keyphrases by training set. Kea was trained using groundwater management texts, with the keywords for each text picked manually or by using the authors suggested keywords. In addition Kea was trained on computer science texts using author keywords. The words given are those picked by Kea when the final document (Kreye et al., 1998) was analysed. The number of matches is the importance given to each term by Kea. As can been seen, non-computer science terms are rank very slightly less importantly under the computer science training set and geographical terms appear slightly less often.

Manually picked keywords	**No. Of Kea matches**	**Author keywords**	**No. Of Kea matches**	**Computer science keywords**	**No. Of Kea matches**
glacier	0.65	glacier	0.63	glacier	0.55
sediment	0.65	sediment	0.63	sediment	0.55
sediment and solute	0.65	sediment and solute	0.63	basins	0.31
basins	0.65	suspended sediment	0.3	time series	0.29
solute transfer	0.45	Broggerbreen	0.3	sediment and solute	0.26
fluvial sediment	0.45	Austre Broggerbreen	0.3	seasonal	0.18
glacio fluvial	0.45	meltwaters	0.3	al	0.16
glacio fluvial sediment	0.45	discharge	0.3	suspended sediment	0.11
suspended sediment	0.43	glacier basins	0.3	glacier basins	0.11
Broggerbreen	0.43	solute transfer	0.16	proglacial	0.11

Table 3: Comparing extracted keyphrases by training set. The paper used was Hodson and Ferguson (1999). See description for Table 2 for details.

Kea can match on average between one and two of the five keyphrases chosen by the papers' authors. However, it must choose from many thousands of candidates. Also, it is highly unlikely that even another human would select the same set of phrases as the original authors. There are some circumstances in which words chosen by the author as keywords do not actually appear anywhere in the text, making it impossible for an automated system to match them. In addition, keywords that are returned which are not author keywords often seem useful for locating similar texts.

4.6.3 Comparison

As can be observed from Tables 2, 3 (above), 4 and 5 (below), automatically extracted keyphrases are not always perfect indicators of a document's content: they often pick words which are far too common to be considered real indicators of a text's subject

area (for example, “course”, “retrieval”, “moisture”). However, they also often provide many extra words that *are* representative of the text's content with which further similar texts can be found (for example, “lysimeter”, “meltwaters”, “natural vegetation” and “emissions”).

Two differences between Extractor and Kea are that Extractor does not pick proper nouns as keywords (i.e. any word which only ever begins with a capital letter), and it does not allow stopwords[16] in the middle of words (so, “sediment and solute” and “retrieval of soil” are not selected as keyphrases). The relative merits of Extractor’s simpler keyword sets in searching for additional resources have to be balanced against Kea’s ability to pick up common academic phrases and place names.

Author keywords	**Kea**	**Extractor**
SAR	**soil moisture**	**soil moisture**
soil moisture	soil moisture content	SMC
evapotranspiration	moisture	measurements
	lysimeter	**SAR**
	SAR	lysimeter
	retrieval	SAR data
	retrieval of soil	natural vegetation
	naturally vegetated	
	probe	
	weighing lysimeter	

Table 4: Comparing keyphrase extraction by Kea and Extractor. Words listed in the order selected by the different methods. Underlined terms are common to the author’s selections. From the paper: Fox et al. (1997) “Retrieval Of Soil Moisture Content From Naturally Vegetated Upland Areas Using ERS-1/2 Synthetic Aperture Radar”.

[16] Stopwords are words such as 'of', 'the', 'and', 'to', 'for', which are considered too frequent in the English language to function as reliable indicators of text type. In the case of Extractor they are picked from the top level words in a word frequency list based on the Brown corpus (a corpus of 1,014,312 words of running English text).

Author keywords	**Kea**	**Extractor**
glacier hydrology	glacier	glacier
suspended sediments	sediment	**suspended sediment**
solutes	sediment and solute	glacier basins
proglacial time series	**suspended sediment**	**solute**
Arctic glaciers	Broggerbreen	discharge
	Austre Broggerbreen	regression models
	meltwaters	solute transfer
	discharge	
	glacier basins	
	solute transfer	

Table 5: Comparing keyphrase extraction by Kea and Extractor. Underlined terms are common to the author's selections. From the paper: Hodson (1999) "Glacio-fluvial sediment and solute transfer in high Arctic basins: examples from Svalbard"

4.7. Hierarchical classification of texts

In this section we will examine how a hierarchical system that already exists can be utilised during the text classification process. As an example, we use the Tellus system (http://www.tellus.ac.uk/), which is a portal for Higher Educational material in Geography, Geology and the Earth Sciences.

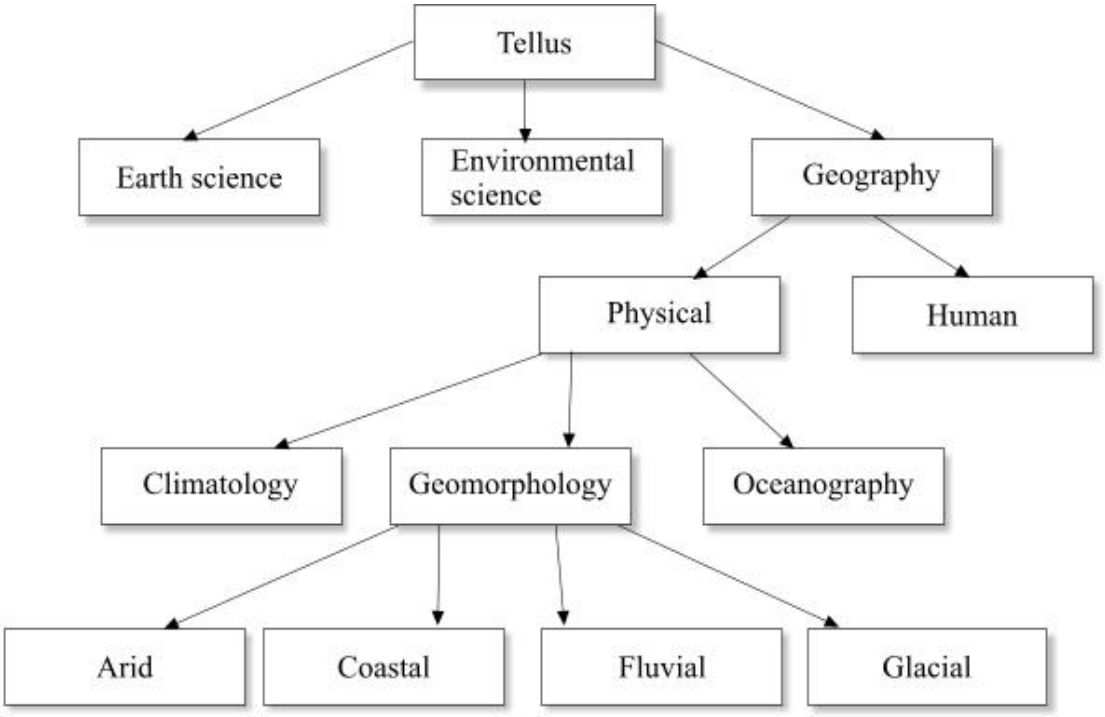

Figure 8: Topic hierarchy of Tellus directory.

By examining pages stored in a pre-existing hierarchy for distinguishing features, such as keywords or structure, common characteristics for specific nodes can be identified. If all the child nodes of a particular node share some features then these features could be taken as characteristic of some higher, parent, node. For example, in the hierarchy shown in Figure 8, if Arid, Coastal, Fluvial and Glacial pages all shared some keywords, then these would be removed from each group and instead assigned to Geomorphology.

Using a hierarchy in this way has a number of advantages. Firstly, common features, including stopwords, will naturally rise to the root where they will not participate in any rankings. These features would be useful for identifying, for example,

'Geography' text but not for distinguishing between different branches of Geography.

Secondly, words that are important for making fine distinctions among categories farther down in the category hierarchy but are ambiguous at higher levels in theory should participate only in places where they can help.

Each node in the hierarchy has a relatively small number of keyphrases that distinguish between the two categories either side of the node. These keyphrases could either be set by experts in the fields in question or could be created by automated keyphrase extraction themselves.

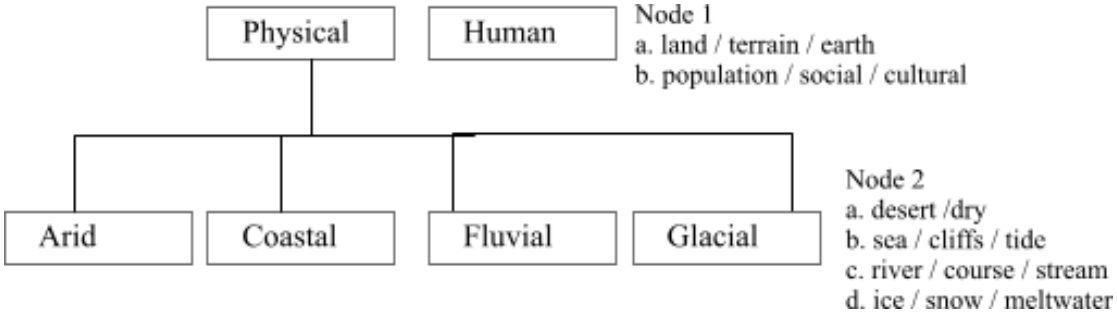

Figure 9: Sample keywords at different levels of Tellus topic hierarchy.

A structured topic hierarchy enables the complex problem of text classification to be broken up into manageable size pieces. Based on the hypothesis that topics which are close to each other in the hierarchy typically have a lot more in common with each other than topics that are far apart, the seemingly large classification task can be divided into a set of smaller classification problems, corresponding to splits in the hierarchy.

This can be seen in Figure 8, where those words that distinguish between subject categories are clearly very different at Nodes 1 and 2 To include the keywords at Node 2 at Node 1 would be wasteful, as many texts at Node 1 will not contain any of these words (the subject area at this point is very wide indeed). Conversely, to include the keywords found at Node 1 at Node 2 is similarly wasteful as the categories on either side of Node 2 are both as likely as each other to contain the more general

keywords which have a more defining role to play closer to the root.

This placing of keywords at each node point in the topic hierarchy makes each sub-task within the overall text classification task much simpler, since at each node in the hierarchy the classifier need only distinguish between a small number of categories. It has been claimed that a key problem in text classification work is the large number of features that are necessary in order to efficiently split texts into separate categories, which can lead to the task becoming unreasonably slow. It would appear that the key to performing quick and robust text classification is the integration of feature selection into a hierarchical structure.

Using a hierarchy can have a positive impact on the categorisation task. Precision and recall are increased and the processing time is substantially reduced.

5. CONVERTING HTML, PDF, AND PS FILE FORMATS FOR TEXTUAL ANALYSIS

Many academic research papers on the web are available in PDF or Postscript format, while some are in Microsoft propriety format. Relatively few are available in HTML. Any automated text location system would have to be capable of pulling plain text from these files. Fortunately, freely available tools, such as GhostView, can do this provided, as is sometimes the case, the PDF files have not been encrypted. In addition, the search engine company Google now offers an API based on the Simple Object Access Protocol (SOAP) and Web Service Definition Language (WSDL). This allows developers to build Google's search facility directly into applications using the Java, Perl, or Visual Studio .NET programming platforms, and therefore gain access to documents translated from PDF and Microsoft formats into Hypertext Markup Language (HTML) (Google, 2002).

6. EXISTING AUTOMATED PORTALS

The creation of topic specific web portals has exploded in recent years as the increasing growth of the World Wide Web has made the location of material which is of interest to the individual user an ever more difficult task. However, the majority of these subject specific portals employ full time staff to read through submitted texts

and categorise them accordingly. The obvious disadvantages to this approach are that material that is erased from the web, or changed, is not updated unless the original submitter contacts the portal staff to inform them, and new material is added relatively slowly, at great expense. However, automated subject specific portals do exist (perhaps unsurprisingly these often cover the field of computer science), and two of the largest are discussed in more detail below.

To date, however, the vast majority of portals are not automated and most provide reasons for why a manned approach has been chosen. Many include the words 'quality' and 'hand-picked' or 'reviewed by experts' side by side, suggesting manual selection and classification were necessary to achieve a quality directory. However, on the other hand, many such directories are not updated very regularly and some smaller ones would appear not to have been updated for more than a year. It seems likely such projects have collapsed due to the lack of resources and manpower required to keep such a project alive.

A typical example of this form of justification can be found on three existing portals detailed below.

iLoveLanguages is a comprehensive catalog of language-related Internet resources. The more than 2000 links at iLoveLanguages have been **hand-reviewed** to bring you **the best language links** the Web has to offer.

http://www.ilovelanguages.com/

This resource list, by no means comprehensive (hundreds of fresh WWW pages are appearing each month), aspires to lend starting points for mining the WWW for foreign language/culture specific resources. This is a **"quality-only" index**. In other words, we have sought to **include only the best** of the foreign language ("foreign" for native speakers of English) Web sites out of the many that exist.

http://www.itp.berkeley.edu/~thorne/HumanResources.html

Resources being added to the Database are **selected, catalogued, classified and subject-indexed by experts** to ensure that **only current, high-quality or useful resources are included.**

http://www.eevl.ac.uk/

Automated portals are possible, however, as can be seen from two thriving working

examples, Cora and ResearchIndex (formerly CiteSeer).

6.1 Cora

Figure 10: screenshot of [http://cora.whizbang.com/about.html -](http://cora.whizbang.com/about.html) the Cora directory.

Cora (Figure 10) is a special-purpose directory of computer science research papers whose creation has been led by Andrew McCallum at Carnegie Mellon University. Cora is the result of McCallum's ongoing research into the field of Machine Learning applied to document classification, information extraction, clustering and crawling (for example, McCallum *et al.*, 2000).

Cora has a Yahoo! style topic hierarchy which contains approximately 75 leaves. Within these leaves there are more than 50,000 academic papers that have, in the

majority, been collected automatically (although there is an additional facility for individuals to add their papers directly to the index).

According to the authors, the construction of Cora was greatly automated by taking advantage of Artificial Intelligence and Machine Learning techniques.

The papers are found by performing topic-directed crawling, using reinforcement learning. The starting point for the crawling was approximately 100 academic computer science departments and industry labs. Papers are then automatically categorised into the topic hierarchy using probabilistic techniques.

In addition to links, Cora provides both citation references (noting both papers which are cited in the current paper and, in turn, those which cite the current paper) and papers' titles and authors which are automatically extracted from the texts using hidden Markov models.

The authors of Cora claim that it is capable of placing documents with 66% accuracy. This figure may not seem particularly high, but it is, on the contrary, very impressive when it is realised that this figure is approaching human agreement levels (when faced, for example, with a text discussing the use of a GIS to analyse volcano eruptions in Sicily, some experts may place the paper under GIS whilst others might be more keen to place it under volcanology). The problem of accurately classifying texts is particularly difficult in the fields, like geography, where both external and internal boundaries in the field often seem fuzzy. For example, deciding whether a text is a sociology text or a social geography text, whether it is an economic text or an economic geography text could prove difficult, even for an expert. One point that must be borne in mind is that a machine can never be expected to accurately classify texts that two human subjects may not necessarily agree on.

Cora's authors make the point that directories such as Yahoo! hire full time staff to manually categorise webpages into their hierarchies, Cora does the same thing automatically, without the need for human effort. They claim that their hierarchy was created in one hour and that the few keywords needed at each node point in their 75 leaf directory took 90 minutes to select. Whilst this process may take longer for diverse fields such as geography, it can be seen that this process has clear advantages.

One current technical limitation with the engine behind Cora, as it stands, is that it can only handle Postscript files. Whilst this is less of a problem for academic papers, a significant percentage of which are still available in this format in many subjects, it would be next to useless in finding teaching material.

6.2 Research Index

ResearchIndex (formerly known as CiteSeer) is an automated directory that currently indexes more than 300,000 pages of postscript and PDF computer science research articles found on the Web. In addition, it provides autonomous citation indexing and automatic notification of new citations and new papers when they match a user profile. The portal locates related documents using citation and word based measures in a continuous update cycle that runs 24 hours a day.

The full source code of ResearchIndex is available free of charge for non-commercial use. Details of the availability of this and all of the other tools and projects mentioned in this report can be found in Appendix A.

7. DISCUSSION

This investigation has outlined and explored the current potential of automated textual analysis tools, and has laid out the basic methodologies and component chains involved in constructing an automated text location and classification system.

It is undoubtedly true that any such system will need to be run in conjunction with a human editor / editors, but that the work load of that editor would be considerably less than would be required with no automated support.

Before implementing any system, a solid resource of sample hand selected and classified texts would need to be assembled as training and validation material for the automated systems to work from.

A tree of topic categories would need to be decided upon as a starting point. Ideally, this would be hierarchical, containing 4 or more levels.

For each proposed category, at least 20 texts would need to be provided, preferably

more.

The sample training texts would have to include a good number (100+) of what was considered representative of ‘educational’ or ‘academic’ material if any form of genre analysis was to be undertaken successfully.

With the above in place, the working system can be constructed along these lines:

1) A set of keywords is produced for all of the pages in the sample set.

2) 80% of the available texts are used to train a keyword extraction tool which is then tested on the remaining 20%.

3) A list of keywords is compiled for each required category; in addition, keyword lists are assembled for each branch node in the classification hierarchy.

4) A classifier suitable for use with a focused crawler is built using the available keywords.

5) Focused crawlers are periodically set off, taking pages already in the system as starting points for the crawl. In this way all new submissions would automatically be examined for pointers to other relevant pages.

6) The catalogue is kept current as modified or removed pages are spotted each time the crawl takes place.

7) Pages that provided the best starting points for crawls are presented to users as good resource pages for research on specific topics.

The system, once set up, would work with minimal support.

The single area that requires most additional research is the identification and classification of educational and teaching material. However, before any research could commence a large and well defined corpus of training material would be required. It is hoped that such texts could be identified using educational keyphrases in conjunction with surface and structural cues.

8. THE FUTURE

There are two broad trends that will make the job of collating web-based resources easier in the future: metadata and resource linking.

Metadata describes resources, that is, it is data about data. For example, a webpage may be marked up as containing 'educational materials'. Plainly as resource and data volumes increase, the necessity for metadata markup will become more apparent. As more people provide metadata, hopefully the searching of the web for academic and educational materials will also become a great deal easier.

There is an increasing trend in Internet based resources for people to mark up metadata using the eXtensible Markup Language (XML)[17]. XML is a flexible language for writing your own HTML-like markup tags, unseen by the majority of users but present in the resources they describe. However, because of the inherent flexibility of XML, there are now several disparate initiatives to provide metadata standards covering the description of academic and educational resources.

The Dublin Core standard covers the metadata tagging of resources in very general terms suitable for most academic materials. While not an XML standard as such, it provides fields that can be turned into XML ('author', 'description', etc.). While Dublin Core is entirely suitable for research materials, the educational community need a more detailed set of metadata fields ('audience education level', 'cost', and 'passwords', for example). Because of this there have been a number of suggested XML-based alternatives. While a Dublin Core Educational group[18] does exist, the initiative fast gaining acceptance as the standard is the IMS Global Learning Consortium's educational metadata specifications[19] (IMS were previously Instructional Management Systems).

[17] http://www.w3.org/xml/
[18] http://dublincore.org/groups/education/
[19] http://www.imsglobal.org/ IMS is backed by the UK's Joint Informations Systems Committee (JISC) and is so widely covered by the Centre for Educational Technology Interoperability Standards (CETIS) as to be the *de facto* standard.

The IMS standards do not simply cover marking up course content. They also cover the linking of resources. For example, there are metadata standards for compiling course descriptions, content, and exams into a single resource and marking up student profiles for use with them. Their ultimate vision is to provide the means by which, for example, a student wanting a degree in geography with economics could have a bespoke course automatically made for them and downloaded to their PC without necessarily going through a traditional educational institution. If this vision seems distant, then it should be noted that most Virtual Learning Environments (like, for example, Leeds' Nathan Boddington building[20], Blackboard[21] or Questionmark[22]) have the ability to package their materials up as IMS compliant resources. Plainly such advances will both advantage university departments wishing to ease the workloads on their staff and place them under considerable competitive stress. Initiatives to harvest IMS metadata resource descriptions for search databases and other types of storage are already underway as part of the Open Archives Initiative (OAI)[23].

One of the most obvious difficulties with metadata, however, is that different people could mark up the same resource in different ways. How do you maintain consistent descriptions of what a resource is about and what it is? How do you describe a 'lecture': is there a difference between a lecture that includes practical exercises and a workshop containing some periods of spoken instruction? Such problems are being addressed by a project currently underway that has a much wider remit than simply searching for educational or research resources: the Semantic Web.

The Semantic Web[24] was outlined by Tim Berners-Lee and his colleagues at the W3 Consortium (see, for example, Berners-Lee *et al.*, 2001) as the ultimate extension and fruition of the web. It aims to provide a structure under which computers can search for, and use, information with an understanding of what it refers to. The current architecture for the project involves two main components, the Resource Description Framework (RDF)[25] and the Web Ontology markup specifications[26]. The RDF

[20] http://www.fldu.leeds.ac.uk/bodingtoncommon.html
[21] http://www.blackboard.com/
[22] http://www.questionmark.com/
[23] http://www.openarchives.org/
[24] http://www.w3.org/2001/sw/
[25] http://www.w3.org/rdf/

provides the necessary tags for saying which metadata standard you are using (thereby negating the need to choose IMS over Dublin Core – you can actually use both or either under the RDF), while the Web Ontology markup languages (which are still stabilizing) give developers a framework in which they can embed the context and meaning of their metadata. For example, it is possible to define what a lecture is, and how it relates to common terms. Users searching for a resource can then tell what your metadata term 'lecture' means and compare it with what others supply. Plainly a lecturer does not have to do this – such descriptions will be defined at a community or international level, and the resource provider will just have to link to the standard descriptions to make their resource available. The ultimate aim of the Semantic Web is not simply to make search results more relevant but to contextualise the knowledge on the web, leading the way for the acquisition of knowledge by language-based artificial intelligence systems.

Plainly these are complex specifications, and one would imagine few academics have time to develop resources, let alone make them available under a metadata standard for the uses of artificial intelligences. However, all of the above initiatives are backed by large corporate groups who intend to provide both resource development and distribution software, and resources of their own in direct competition with the academic sector over the coming decade/s. For this reason alone, Academia would do well to pay attention. The advantages for academics from these developments will hopefully be more flexibility in the audiences they reach, and an enhanced ability to find information and resources of use in their work.

[26] http://www.w3.org/2001/sw/WebOnt/

9. REFERENCES

Barker K. and Cornacchia N. (2000) Using noun phrase heads to extract document keyphrases *Advances in Artificial Intelligence, Proceedings, Lecture Notes in Artificial Intelligence, Vl. http://www.cs.utexas.edu/users/kbarker/papers/BarkerCornacchia.pdf* (accessed 8 January 2003)

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. *Scientific American.com http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21* (accessed 8 January 2003)

Biber, D, Conrad, S., & Reppen, R. (1998) *Corpus Linguistics* Cambridge University Press, Cambridge. pp.320.

Chakrabarti S., van den Berg M. and Dom B. (1999) Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery *Paper given at the 8th World Wide Web Conference, Toronto, Canada, 1999 http://www8.org/w8-papers/5a-search-query/crawling/* (accessed 8 January 2003)

Google (2002) Google Web APIs. *http://www.google.com/apis/* (accessed 8 January 2003)

Fischetti, M. V., Sano, N., Laux, S. E., and Natori, K. (1996) Full-band-structure theory of high-field transport and impact ionization of electrons and holes in Ge, Si, and GaAs *IEEE Trans. Semicond. Technol. Modeling and Simulation, special issue: Proceedings of the 1996 International Conference of Semiconductor Processes and Devices (SISPAD'96), (Tokyo, Japan, 1996) http://www.jtcad.tec.ufl.edu/archive.html* (accessed 8 January 2003)

Frank, E., Paynter G.W., Witten, I.H., Gutwin, C. and Nevill-Manning, C.G. (1999) Domain-Specific Keyphrase Extraction *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence http://www.cs.waikato.ac.nz/~eibe/pubs/Z507.ps.gz* (accessed 8 January 2003)

Fox, N.I, Saich, P.J., Collier, C.G. & Miller, R.J. (1997) Retrieval of Soil Moisture Content from Naturally Vegetated Upland Areas using ERS-1/2 Synthetic Aperture

Radar Data *Proc. 3rd ERS Symposium: Space at the Service of our Environment, Florence 1997, ESA SP-414.* http://earth.esa.int/symposia/papers/fox/127c.htm (accessed 8 January 2003)

Hodson, A.J., and Ferguson, R.I., (1999) Fluvial suspended sediment transport from cold and warm-based glaciers in Svalbard, *Earth Surface Processes and Landforms*, 24, 11, 957-974.

Hodson, A.J. (1999) Glacio-fluvial sediment and solute transfer in high Arctic basins: examples from Svalbard. *Glacial Geology and Geomorphology*, rp10/1999 *http://ggg.qub.ac.uk/papers/full/1999/rp101999/rp10.html* (accessed 8 January 2003)

Hunt, W.B., (1996) *Getting to War: Predicting International Conflict With Mass Media Indicators* University of Michigan Press. pp.264.

Kreye, R., Ronneseth, K. and Wei, M. (1998) *An Aquifer Classification System For Groundwater Management In British Columbia*, Ministry Of Environment, Lands And Parks, Water Management Division, Hydrology Branch, Province Of British Columbia *http://Wlapwww.Gov.Bc.Ca/Wat/Aquifers/Aq_Classification/Aq_Class.Html* (accessed 8 January 2003)

Iwanska, L.M, and Shapiro, S.C. (2000) *Natural Language Processing and Knowledge Representation.* AAAI Press / MIT Press. pp.350

ISO (1999) *ISO 12620 Computer applications in terminology -- Data categories*. pp.71

Kessler B., Nunberg G. and Schütze, H. (1997) Automatic Detection of Text Genre. *ACL/EACL 1997 http://acl.ldc.upenn.edu/P/P97/P97-1005.pdf* (accessed 8 January 2003)

Krulwich, B. and Burkey, C. (1996) *Learning user information interests through the extraction of semantically significant phrases* Working Notes of the 1996 AAAI Spring Symposium on Machine Learning in Information Access, March 1996. *http://www2.parc.com/istl/projects/mlia/mlia-papers.shtml* (accessed 8 January 2003)
McCallum, A. K., Nigam, K., Rennie, J., Seymore, K. (2000) Automating the construction of Internet portals with machine learning *Information Retrieval*, 3, 2, 127-163.

Notess, G. R, (2002) Search Engines Statistics: Database Overlap *http://www.searchengineshowdown.com/stats/overlap.shtml* (accessed 4 July 2002)

Oxnard, L. (in prep) *The Automated Analysis of Catalan Texts* Unpublished University of Sheffield Ph.D. thesis.

Turney P D., (2000) Learning Algorithms for Keyphrase Extraction *Information Retrieval* 2, 4, 303-336.

Witten I. H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G. (1999) KEA: Practical Automatic Keyphrase Extraction *Association for Computing Machinery International Conference on Digital Libraries (ACM DL), 1999. http://www.nzdl.org/Kea/Nevill-et-al-1999-DL99-poster.pdf* (accessed 4 July 2002)

APPENDIX A

Software products and tools of interest.

Product	What it is	Web page	Availability
Kea	Keyphrase extractor	http://www.nzdl.org/Kea	GNU public license
Extractor	Keyphrase extractor	http://extractor.iit.nrc.ca/	Licensed (research option)
Citeseer (ResearchIndex)	Autonomous citation index builder	http://www.neci.nec.com/~lawrence/researchindex.html	Full source code available for non-commercial use
GhostView	Plain text converter	http://www.cs.wisc.edu/~ghost/gsview/	GNU public license
QTAG	Part of Speech tagger	http://www.clg.bham.ac.uk/QTAG/	Available for research purposes
WebSPHINX	Web crawler	http://www.cs.cmu.edu/~rcm/websphinx/	GNU public license