AUTOMATIC IDENTIFICATION AND ORGANIZATION OF INDEX TERMS FOR INTERACTIVE BROWSING

Nina Wacholder
Columbia University
New York, NY
nina@cs.columbia.edu

David K. Evans
Columbia University
New York, NY
devans@cs.columbia.edu

Judith L. Klavans
Columbia University
New York, NY
klavans@cs.columbia.edu

ABSTRACT

Indexes -- structured lists of terms that provide access to document content -- have been around since before the invention of printing [31]. But most text content in digital libraries is not accessible through indexes. In this paper, we consider two questions related to the use of automatically identified index terms in interactive browsing applications: 1) Is the quality and quantity of the terms identified by automatic indexing such that they provide useful access points to text in automatic browsing applications? and 2) Can automatic sorting techniques bring terms together in ways that are useful for users?

The terms that we consider have been identified by LinkIT, a software tool for identifying significant topics in text [16]. Over 90% of the terms identified by LinkIT are coherent and therefore merit inclusion in the dynamic text browser. Terms identified by LinkIT are input to a dynamic text browser, a system that supports interactive navigation of index terms, with hyperlinks to the views of phrases in context and full-text documents. The distinction between phrasal heads (the most important words in a coherent term) and modifiers serves as the basis for a hierarchical organization of terms. This linguistically motivated structure helps users to efficiently browsing and disambiguate terms. We conclude that the approach to information access discussed in this paper is very promising, and also that there is much room for further research. In the meantime, this research is a contribution to the establishment of a sound foundation for assessing the usability of terms in phrase browsing applications.

Keywords

Indexing, phrases, natural language processing, browsing, genre.

1. OVERVIEW

Indexes are useful for information seekers because they:

- support browsing, a basic mode of human information seeking [32].
- provide information seekers with a valid list of terms, instead of requiring users to invent the terms on their own. Identifying index terms has been shown to be one of the hardest parts of the search process, e.g., [17].
- are organized in ways that bring related information together [31].

But indexes are not generally available for digital libraries. The manual creation of an index is a time consuming task that requires a considerable investment of human intelligence [31]. Individuals and institutions simply do not have the resources to create expert indexes for digital resources.

However, automatically generated indexes have been legitimately criticized by criticized by information professionals such as Mulvany 1994 [31]. Indexes created by computers systems are different than those compiled by human beings. A certain number of automatically identified index terms inevitably contain errors that look downright foolish to human eyes. Indexes consisting of automatically identified terms have been criticized by grounds that they constitute indiscriminate lists, rather than synthesized and structured representation of content. And because computer systems do not understand the terms they extract, they cannot record terms with the consistency expected of indexes created by human beings.

Nevertheless, the research approach that we take in this paper emphasizes fully automatic identification and organization of index terms that actually occur in the text. We have adopted this approach for several reasons:

1. **Human indexers simply cannot keep up with the volume of new text being produced.** This is a particularly pressing problem for publications such as daily newspapers because they are under particular pressure to rapidly create useful indexes for large amounts of text.
2. **New names and terms are constantly being invented and/or published.** For example, new companies are formed (e.g., *Verizon Communications Inc.*); people's names appear in the news for the first time (e.g., it is unlikely that *Elian Gonzalez*' name was in a newspaper before November 25, 1999); and new product names are constantly being invented (e.g., *Handspring's Visor PDA*). These terms frequently

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Conference '00, Month 1-2, 2000, City, State.
Copyright 2000 ACM 1-58113-000-0/00/0000…$5.00.

appear in print some type before they appear in an authoritative reference source.

3. **Manually created external resources are not available for every corpus.** Systems that fundamentally depend on manually created resources such as controlled vocabularies, semantic ontologies, or the availability of manually annotated text usually cannot be readily adopted to corpora for which these resources do not exist.

4. **Automatically identified index terms are useful in other digital library applications.** Examples are information retrieval, document summarization and classification [43], [2].

In this paper, we describe a method for creating a dynamic text browser, a user-centered system for browsing and navigating index terms. The focus of our work is on the usability of the automatically identified index terms and on the organization of these terms in a ways that reduce the number of terms that users need to browse, while retaining context that helps to disambiguate the terms.

The input to Intell-Index, our dynamic text browser, is the output of a system called LinkIT that automatically identifies significant topics in full text documents. LinkIT efficiently identifies noun phrases in full-text documents in any domain or genre [16], [15]. LinkIT also identifies the head of each noun phrase and creates pointers from each noun phrase head to all expansions that occur in the corpus. The head of a noun phrase is the noun that is semantically and syntactically the most important element in the phrase. For example, *filter* is the head of the noun phrases *coffee filter*, *oil filter*, and *smut filter*. The dynamic text browser supports hierarchical navigation of index terms by heads or by expanded phrases. In addition, Intell-Index allows the user to search the index in order to identify subsets of related terms based on criteria such as frequency of a phrase in a document, or whether the phrase is a proper name. The dynamic text browser thereby supports a mode of navigation of terms that takes advantage of the computer's ability to rapidly process large amounts of text and the human ability to use world knowledge and context to actually understand meaning of terms.

We know of no other work that addresses the specific question of how to assess the usability of automatically identified terms in browsing applications, so we have chosen to focus on three criteria for assessing the usability of the index terms in the dynamic text browser: **quality of index terms**, **thoroughness of coverage of document content** and **sortability of index terms**.

- **Quality of index terms.** Because computer systems are unable to identify terms with human reliability or consistency, they inevitably generate some number of junk terms that humans readily recognize as incoherent. We consider a very basic question: are automatically identified terms sufficiently coherent to be useful as access points to document content. To answer this question for the LinkIT output, we randomly selected .025% of the terms identified in a 250MB corpus and evaluated them with respect to their coherence. Our study showed that over 90% of the terms are coherent. Cowie et Lehnert 1996 [7] observe that 90% precision in information extraction is probably satisfactory for every day use of results; this assessment is relevant here because the terms are processed by people, who can fairly readily ignore the junk if they expect to encounter it.

- T**horoughness of coverage of document content.** Because computer systems are more thorough and less discriminating, they typically identify many more terms than a human indexer would for the same amount of material. For example, LinkIT identifies about 500,000 non-unique terms for 12.27 MB of text. We address the issue of quantity by considering the number of terms that LinkIT identifies, as related to size of the original text from which they were extracted. This provides a basis for future comparison of the number of terms identified in different corpora and by different techniques.

- **Sortability of index terms.** Because electronic presentation supports interactive filtering and sorting of index terms, the actual number of index terms is less important than the availability of useful ways to bring together useful subsets of terms. In this paper, we show that head sorting, a method for sorting index terms discussed in Wacholder 1998 [38], is a linguistically motivated way to sort index terms in ways that provide useful views of single documents and of collections of documents.

This work contributes to our understanding of what constitutes useful terms for browsing and toward the development of effective techniques for filtering and organizing these terms. This reduces the number of terms that the information seeker needs to scan, while maximizing the information that the user can obtain from the list of terms.

There is an emerging body of related work on development of interactive systems to support phrase browsing (e.g., Anick and Vaithyanathan 1997 [2], Gutwin et al. [19], Nevill-Manning et al. 1997 [32], Godby and Reighart 1998 [18]). The criteria that we identify for assessing our own system, term quality, thoroughness of coverage and sortability can be used in future work to determine what properties of this type of system are most useful.

We will discuss of term quality and thoroughness of coverage of document content in Section 3. Sortability of index terms is discussed in Section 4. But before turning to these issues, we present Intell-Index, our dynamic text browser.

2. Intell-Index, a dynamic text browser.

One of the fundamental advantages of an electronic browsing environment relative to a printed one is that the electronic environment readily allows a single item to be viewed in many contexts. To explore the promise of dynamic text browsers for browsing index terms and linking from index terms to full-text documents, we have implemented a prototype dynamic text browser, called Intell-Index, which allows users to interactively sort and browse terms.

Figure 1 on p.9 shows the Intell-Index opening screen. The user has the option of either browsing all of the index terms identified in the corpus or specifying a search string that index terms should match. Figure 2 on p.9 shows the beginning of the alphebetized browsing results for the specified corpus. The user may click on a term to view the context in which the term is used; these contexts are sorted by document and ranked by normalized frequency in the document. This is a version of KWIC (keyword in context) that we call ITIC (index term in context). Finally, if the set of

ITICs for a document suggest that the document is relevant, the user may choose to view the entire document.

However, the large number of terms listed in indexes makes it important to offer alternatives to browsing the complete list of index terms identified for a corpus. Information seekers can view a subset of the complete list by specifying a search string. Search criteria implemented in Intell-Index include:

- **case matching**: whether or not the terms returned must match the case of the user-specified search string. This facility allows the user to view only proper names (with a capitalized last word), only common noun phrases, or both. This is an especially useful facility for controlling terms that the system returns. For example, specifying that the *a* in *act* be capitalized in a collection of social science or political articles is likely to return a list of laws with the word *Act* in their title; this is much more specific than an indiscriminate search for the string *act*, regardless of capitalization.

- **string matching**: whether or not the search string must occur as a single word. This facility lets the user control the breadth of the search: a search for a given string as a word will usually return fewer results than a search for the string as a substring of larger words. For very common words, the substring option is likely to produce more terms than the user wants; for example, a search for the initial substring *act* will return *act(s)*, *action(s), activity, activities*, *actor(s), actual*, *actuary, actuaries* etc, but sometimes is very convenient because it will return different morphological forms of a word, e.g., *activit* will return occurrences of *activity* and *activities*. The word match option is particularly useful for looking for named entities.

- **location of search string in phrase**: whether the search string must occur in the head of the simplex noun phrase, the modifier (i.e., words other than the head), or anywhere in the term. By specifying that the search string must occur in the head of the index term, as with *worker*, the user is likely to obtain references to kinds of workers, such as *asbestos workers*, *hospital workers*, *union workers* and so forth. By specifying that the search term must occur as a modifier, the user is likely to obtain references to topics discussed specifically with regard to their impact on workers, as in *workers' rights, worker compensation, worker safety*, *worker bees*.

In addition, the information seeker has options for sorting the terms. For example, the user can ask for terms to be alphabetized from left to right, as is standard. In addition, the user can sort the words by head and in the order in which they occurred in the original document.

Because of the functionality of dynamic text browsers, terms may be useful in the dynamic text browser that are not useful in alphabetical lists of terms. In the next section we assess, qualitatively and quantitatively, the usability of automatically indexed terms in this type of application.

3. Automatically identified index terms

3.1 Quality

The problem of how to determine what index terms merit inclusion in a dynamic text browsing application is a difficult one. The standard information retrieval metrics of precision and recall do not apply to this task because indexes are designed to satisfy multiple information needs. In information retrieval, precision is calculated by determining how many retrieved documents satisfy a specific information need. But indexes by design include index terms that are relevant to a variety of information needs. To apply the recall metric to index terms, we would calculate the proportion of good index terms correctly identified by a system relative to the list of all possible good index terms. But we do not know what the list of all possible good index terms should look like. Even comparing an automatically generated list to a human generated list is difficult because human indexers add index entries that do not appear in the text; this would bias the evaluation against an index that only includes terms that actually occur in the text.

In this section we therefore consider a baseline property of index terms: coherence. This is important because any list of automatically identified terms inevitably includes some junk, which inevitably detracts from the usefulness of the index.

To assess the coherence of automatically identified index terms, 583 index terms (.025% of the total) were randomly extracted from the 250 MB corpus and alphabetized. Each term was assigned one of three ratings:

- **coherent** -- a term is both coherent and a noun phrase. arguably a coherent noun phrase. Coherent terms make sense as a distinct unit, even out of context. Examples of coherent terms identified by LinkIT are *sudden current shifts*, *Governor Dukakis*, *terminal-to-host connectivity* and *researchers*.

- **incoherent** – a term is neither a noun phrase nor coherent. Examples of incoherent terms identified by LinkIT are *uncertainty is*, *x ix limit*, and *heated potato then shot*. Most of these problems result from idiosyncratic or non-standard text formatting. Another source of errors is the part-of-speech tagger; for example, if it erroneously identifies a verb as a noun (as in the example *uncertainty is*), the resulting term is incoherent.

- **intermediate** – any term that does not clearly belong in the coherent or incoherent categories. Typically they consist of one or more good noun phrases, along with some junk. In general, they are enough like noun phrases that in some ways they fit patterns of the component noun phrases. One example is *up Microsoft Windows*, which would be a coherent term if it did not include *up*. We include this term because the term is coherent enough to justify inclusion in a list of references to Windows or Microsoft. Another example is *th newsroom*, where *th* is presumably a typographical error for *the*. There are a higher percentage of intermediate terms among proper names than the other two categories; this is because LinkIT has difficulty of deciding where one proper name ends and the next one begins, as in *General Electric Co. MUNICIPALS Forest Reserve District*.

Table 1 shows the ratings by type of term and overall. The percentage of useless terms is 6.5%. This is well under 10%, which puts our results in the realm of being suitable for everyday use according to the Cowie and Lehnert metric mentioned in Section 1.

Table 1: Quality rating of terms, as measured by comprehensibility of terms[1]

	Total	Cohe-rent	Interme-diate	Inco-herent
Number of words	574	475	62	37
% of total words	100%	82.8%	10.9%	6.5%

In a previous study we conducted an experiment in which users were asked to evaluate index terms identified by LinkIT and two other domain-independent methods for identifying index terms in text (Wacholder et al. 2000 [40]). This study showed that when compared to the other two methods by a metric that combines quality of terms and coverage of content, LinkIT was superior to the other two techniques.

These two studies demonstrate that automatically identified terms like those identified by LinkIT are of sufficient quality to be useful in browsing applications. We plan to conduct additional studies that address the issue of the usefulness of these terms; one example is to give subjects indexes with different terms and see how long it takes them to satisfy a specific information need.

3.2 Thoroughness of coverage of document content

Thoroughness of coverage of document content is a standard criterion for evaluation of traditional indexes [20]. In order to establish an initial measure of thoroughness, we evaluate number of terms identified relative to the size of the text.

Table 2 shows the relationship between document size in words and number of noun phrases per document. For example, for the AP corpus, an average document of 476 words typically has about 127 non-unique noun phrases associated with it. In other words, a user who wanted to view the context in which each noun phrase occurred would have to look at 127 contexts. (To allow for differences across corpora, we report on overall statistics and per corpus statistics as appropriate.)

[1] For this study, we eliminated terms that started with non-alphabetic characters.

Table 2: Noun phrases (NPs) per document

Corpus	Avg. Doc Size	Avg. number of NPs/doc
AP	2.99K (476 words)	127
FR	7.70K (1175 words)	338
WSJ	3.23K (487 words)	132
ZIFF	2.96K (461 words)	129

The numbers in Table 2 are important because they vary radically depending on the technique used to identify noun phrases. Noun phrases as they occur in natural language are recursive, that is noun phrases occur within noun phrases. For example, the complex noun phrase *a form of cancer-causing asbestos* actually includes two simplex noun phrases, *a form* and *cancer-causing asbestos*. A system that lists only complex noun phrases would list only one term, a system that lists both simplex and complex noun phrases would list all three phrases, and a system that identifies only simplex noun phrases would list two.

A human indexer readily chooses whichever type of phrase is appropriate for the content, but natural language processing systems cannot do this reliably. Because of the ambiguity of natural language, it is much easier to identify the boundaries of simplex noun than complex ones [38]. We therefore made the decision to focus on simplex noun phrases rather than complex ones for purely practical reasons.

The option of including both complex and simple forms was adopted by Tolle and Chen 2000 [35]. They identify approximately 140 unique noun phrases per abstract for 10 medical abstracts. They do not report the average length in words of abstracts, but a reasonable guess is probably about 250 words per abstract. On this calculation, the relation between the number of noun phrases and the number of words in the text is .56. In contrast, LinkIT identifies about 130 NPs for documents of approximately 475 words, for a ratio of just under 500 words, for a ratio of .27. The index terms represent the content of different units: 140 index terms represents the abstract, which is itself only an abbreviated representation of the document. The 130 terms identified by LinkIT represent the entire text, but our intuition is that it is better to provide coverage of full documents than of abstracts. Experiments to determine which technique is more useful for information seekers are needed

For each full-text corpus, we created one parallel version consisting only of all occurrences of all noun phrases (duplicates not removed) in the corpus, and another parallel version consisting only of heads (duplicates not removed), as shown in Table 3. The numbers in parenthesis are the number of words per document and per corpus for the full-text columns, and the percentage of the full text size for the noun phrase (NP) and head column.

Table 3 Corpus Size

Corpus	Full Text	Non Unique NPs	Unique NPs
AP	12.27 MB (2.0 million words)	7.4 MB (60%)	2.9 MB (23%)
FR	33.88 MB (5.3 million words)	20.7 MB (61%)	5.7 MB (17%)
WSJ	45.59 MB (7.0 million words)	27.3 MB (60%)	10.0 MB (22%)
ZIFF	165.41 MB (26.3 million words)	108.8 MB (66%)	38.7 MB (24%)

The number of noun phrases reflects the number of occurrences (tokens) of NPs and heads of NPs. Interestingly, the percentages are relatively consistent across corpora.

From the point of view of the index, however, the figures shown in Table 3 represent only a first level reduction in the number of candidate index terms: for browsing and indexing, each term need be listed only once. After duplicates have been removed, approximately 1% of the full text remains for heads, and 22% for noun phrases. This suggests that we should use a hierarchical browsing strategy, using the shorter list of heads for initial browsing, and then using the more specific information in the fuller noun phrases when specification is requested. The implications of this are explored in Section 4.

4. Sortability of index terms

Human beings readily use context and world knowledge to interpret information. Structured lists are particularly useful to people because they bring related terms together, either in documents or across documents. In this section, we show some methods for organizing terms that can readily be accomplished automatically, but take too much effort and space to be used in printed indexes for corpora of any size.

One linguistically motivated way for sorting index terms is by head, i.e., by the element that is semantically and syntactically the most important element in a phrase. Index terms in a document, i.e., the noun phrases identified by LinkIT, are sorted by head, the element that is linguistically recognized as semantically and syntactically the most important. The terms are ranked in terms of their significance based on frequency of the head in the document, as described in Wacholder 1998 [38]. After filtering based on significance ranking and other linguistic information, the following topics are identified as most important in a single article extracted from *Wall Street Journal 1988*, available from the Penn Treebank. (Heads of terms are italicized.)

Table 4: **Most significant terms in document**

asbestos *workers*
cancer-causing *asbestos*
cigarette *filters*
researcher(s)
asbestos *fiber*
crocidolite
paper *factory*

This list of phrases (which includes heads that occur above a frequency cutoff of 3 in this document, with content-bearing modifiers, if any) is a list of important concepts representative of the entire document.

Another view of the phrases enabled by head sorting is obtained by linking noun phrases in a document with the same head. A single word noun phrase can be quite ambiguous, especially if it is a frequently-occurring noun like *worker*, *state*, or *act*. Noun phrases grouped by head are likely to refer to the same concept, if not always to the same entity (Yarowsky 1993 [42]), and therefore convey the primary sense of the head as used in the text. For example, in the sentence "Those workers got a pay raise but the other workers did not", the same sense of *worker* is used in both noun phrases even though two different sets of workers are referred to. Table 5 shows how the word *workers* is used as the head of a noun phrase in four different Wall Street Journal articles from the Penn Treebank; determiners such as *a* and *some* have been removed.

Table 5: **Comparison of uses of *worker* as head of noun phrases across articles**

workers … asbestos workers (wsj 0003)
workers … private sector workers … private sector hospital workers ... nonunion workers…private sector union workers (wsj 0319)
workers … private sector workers … United Steelworkers (wsj 0592)
workers … United Auto Workers … hourly production and maintenance workers (wsj0492)

This view distinguishes the type of *worker* referred to in the different articles, thereby providing information that helps rule in certain articles as possibilities and eliminate others. This is because the list of complete uses of the head *worker* provides explicit positive and implicit negative evidence about kinds of workers discussed in the article. For example, since the list for wsj_0003 includes only *workers* and *asbestos workers*, the user can infer that hospital workers or union workers are probably not referred to in this document.

Term context can also be useful if terms are presented in document order. For example, the index terms in Table 6 were extracted automatically by the LinkIT system as part of the process of identification of all noun phrases in a document (Evans 1998 [15]; Evans et al. 2000[16].

Table 6: Topics, in document order, extracted from first sentence of wsj0003

A form
asbestos
Kent cigarette filters
a high percentage
cancer deaths
a group
workers
30 years
researchers

For most people, it is not difficult to guess that this list of terms has been extracted from a discussion about deaths from cancer in workers exposed to asbestos. The information seeker is able to apply common sense and general knowledge of the world to interpret the terms and their possible relation to each other. At least for a short document, a complete list of terms extracted from a document in order can relatively easily be browsed in order to get a sense of the topics discussed in a single document.

The three tables above show just a few of the ways that automatically identified terms are organized and filtered in our dynamic text browser.

In the remainder of this section, we consider how a dynamic text browser which has information about noun phrases and their heads helps facilitate effective browsing by reduce the number of terms that an information seeker needs to look at.

In general, the number of unique noun phrases increases much faster than the number of unique heads – this can be seen by the fall in the ratio of unique heads to noun phrases as the corpus size increases.

Table 7: Number of unique noun phrases(NPs) and heads

Corpus	Unique NPs	Unique Heads	Ratio of Unique Heads to NPs
AP	156798	38232	24%
FR	281931	56555	20%
WSJ	510194	77168	15%
ZIFF	1731940	176639	10%
Total	2490958	254724	10%

Table 7 is interesting for a number of reasons:

1) the variation in ratio of heads to noun phrases per corpus—this may well reflect the diversity of AP and the FR relative to the WSJ and especially Ziff.
2) as one would expect, the ration of heads to the total is smaller for the total than for the average of the individual corpora. This is because the heads are nouns. (No dictionary can list all nouns; this list is constantly growing, but at a slower rate than the possible number of noun phrases).

In general, the vast majority of heads have two or fewer different possible expansions. There is a small number of heads, however, that contain a large number of expansions. For these heads, we could create a hierarchical index that is only displayed when the user requests further information on the particular head. In the data that we examined, on average the heads had about 6.5 expansions, with a standard deviation of 47.3.

Table 8: Average number of head expansions per corpus

Corp	Max	% <= 2	2 < % < 50	% >= 50	Avg	Std. Dev.
AP	557	72.2%	26.6%	1.2%	4.3	13.63
FR	1303	76.9%	21.3%	1.8%	5.5	26.95
WSJ	5343	69.9%	27.8%	2.3%	7.0	46.65
ZIFF	15877	75.9%	21.6%	2.5%	10.5	102.38

The most frequent head in the Ziff corpus, a computer publication, is *system*.

Additionally, these terms have not been filtered; we may be able to greatly narrow the search space if the user can provide us with further information about the type of terms they are interested in. For example, using simple regular expressions, we are able to roughly categorize the terms that we have found into four categories: noun phrases, SNPs that look like proper nouns, SNPs that look like acronyms, and SNPs that start with non-alphabetic characters. It is possible to narrow the index to one of these categories, or exclude some of them from the index.

Table 9: Number of SNPs by category

Corpus	# of SNPs	# of Proper Nouns	# of Acronyms	# of non-alphabetic elements
AP	156798	20787 (13.2%)	2526 (1.61%)	12238 (7.8%)
FR	281931	22194 (7.8%)	5082 (1.80%)	44992 (15.95%)
WSJ	510194	44035 (8.6%)	6295 (1.23%)	63686 (12.48%)
ZIFF	1731940	102615 (5.9%)	38460 (2.22%)	193340 (11.16%)
Total	2490958	189631 (7.6%)	45966 (1.84%)	300373 (12.06%)

For example, over all of the corpora, about 10% of the SNPs start with a non-alphabetic character, which we can exclude if the user is searching for a general term. If we know that the user is searching specifically for a person, then we can use the list of proper nouns as index terms, further narrowing the search space to approximately 10% of the possible terms.

5. CONCLUSION

When we began working on this paper, our goal was simply to assess the quality of the terms automatically identified by LinkIT for use in electronic browsing applications. Through an evaluation of the results of an automatic index term extraction system, we have shown that automatically generated indexes can be useful in a dynamic text-browsing environment such as Intell-Index for enabling access to digital libraries.

We found that natural language processing techniques have reached the point of being able to reliably identify terms that are coherent enough to merit inclusion in a dynamic text browser: over 93% of the index terms extracted for use in the Intell-Index system have been shown to be useful index terms in our study. This number is a baseline; the goal for us and others should be to improve these numbers.

We have also demonstrated how sorting of index terms by head makes it easier to browse index terms. The possibilities for additional sorting and filtering index terms are multiple, and our work suggests that these possibilities are worthy of exploration. Our results have implications for our own work and also for research results with regard to phrase browsers referred to in Section 1.

As we conducted this work, we discovered that there are many unanswered questions about the usability of index terms. In spite of a long history of indexes as an information access tool, there has been relatively little research on indexing usability, an especially important topic vis a vis automatically generated indexes [20][30].

Among them are the following:

1. What properties determine the usability of index terms?
2. How is the usefulness of index terms affected by the browsing environment?
3. From the point of view of representation of document content, what is the optimal relationship between number of index terms and document size?
4. What number of terms can information seekers readily browse? Do these numbers vary depending on the skill and domain knowledge of the user?

Because of the need to develop new methods to improve access to digital libraries, answering questions about index usability is a research priority in the digital library field. This paper makes two contributions: description of a linguistically motivated method for identifying and browsing index terms and establishment of fundamental criteria for measuring the usability of terms in phrase browsing applications.

6. ACKNOWLEDGMENTS

This work has been supported under NSF Grant IRI-97-12069, "Automatic Identification of Significant Topics in Domain Independent Full Text Analysis", PI's: Judith L. Klavans and Nina Wacholder and NSF Grant CDA-97-53054 "Computationally Tractable Methods for Document Analysis", PI: Nina Wacholder.

7. REFERENCES

[1] Aberdeen, J., J. Burger, D. Day, L. Hirschman, and M. Vilain (1995) "Description of the *Alembic* system used for MUC-6". In *Proceedings of MUC-6*, Morgan Kaufmann. Also, Alembic Workbench, http://www.mitre.org/resources/centers/advanced_info/g04h/workbench.html.

[2] Anick, Peter and Shivakumar Vaithyanathan (1997) "Exploiting clustering and phrases for context-based information retrieval", *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR *'97)*, pp.314-323.

[3] Baeza-Yates, Ricardo and Berthier Ribeiro-Netto (1999) *Modern Information Retrieval*, ACM Press, New York.

[4] Bagga, Amit and Breck Baldwin (1998). "Entity based cross-document coreferencing using the vector space model, *Proceeding of the 36th Annual Meeting of the Association forComputational Linguistics and the 17th International Conference on Computational Linguistics*, pp.79-85.

[5] Bikel, D., S. Miller, R. Schwartz, and R. Weischedel (1997) Nymble: a High-Performance Learning Name-finder", *Proceedings of the Fifth conference on Applied Natural Language Processing*, 1997.

[6] Boguraev, Branimir and Kennedy, Christopher (1998) "Applications of term identification Terminology: domain description and content characterisation", *Natural Language Engineering* 1(1):1-28.

[7] Cowie, Jim and Wendy Lehnert (1996) "Information extraction", *Communications of the ACM*, 39(1):80-91.

[8] Church, Kenneth W. (1998) "A stochastic parts program and noun phrase parser for unrestricted text", *Proceedings of the Second Conference on Applied Natural Language Processing*, 136-143.

[9] Dagan, Ido and Ken Church (1994) *Termight*: Identifying and translating technical terminology, *Proceedings of ANLP '94*, Applied Natural Language Processing Conference, Association for Computational Linguistics, 1994.

[10] Damereau, Fred J. (1993) "Generating and evaluating domain-oriented multi-word terms from texts", *Information Processing and Management* 29(4):433-447.

[11] DARPA (1998) *Proceedings of the Seventh Message Understanding Conference* (MUC-7). Morgan Kaufmann, 1998.

[12] DARPA (1995) *Proceedings of the Sixth Message Understanding Conference* (MUC-6). Morgan Kaufmann, 1995.

[13] Edmundson, H.P. and Wyllys, W. (1961) "Automatic abstracting and indexing--survey and recommendations", *Communications of the ACM*, 4:226-234.

[14] Evans, David A. and Chengxiang Zhai (1996) "Noun-phrase analysis in unrestricted text for information retrieval", *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp.17-24. 24-27 June 1996, University of California, Santa Cruz, California, Morgan Kaufmann Publishers.

[15] Evans, David K. (1998) LinkIT Documentation, Columbia University Department of Computer Science Report. Available at <http://www.cs.columbia.edu/~devans/papers/LinkITTechDoc/ >

[16] Evans, David K., Klavans, Judith, and Wacholder, Nina (2000) "Document processing with LinkIT", *Proceedings of the RIAO Conference,* Paris, France.

[17] Furnas, George, Thomas K. Landauer, Louis Gomez and Susan Dumais (1987) "The vocabulary problem in human-system communication", *Communications of the ACM* 30:964-971.

[18] Godby, Carol Jean and Ray Reighart (1998) "Using machine-readable text as a source of novel vocabulary to update the Dewey Decimal Classification", presented at the SIG-CR Workshop, ASIS, < http://orc.rsch.oclc.org:5061/papers/sigcr98.html >.

[19] Gutwin, Carl, Gordon Paynter, Ian Witten, Craig Nevill-Manning and Eibe Franke (1999) "Improving browsing in digital libraries with keyphrase indexes*", Decision Support Systems* 27(1-2):81-104.

[20] Hert, Carol A., Elin K. Jacob and Patrick Dawson (2000) "A usability assessment of online indexing structures in the networked environment", *Journal of the American Society for Information Science* 51(11):971-988.

[21] Hatzivassiloglou, Vasileios, Luis Gravano, and Ankineedu Maganti (2000) "An investigation of linguistic features and clustering algorithms for topical document clustering," *Proceedings of Information Retrieval* (SIGIR'00), pp.224-231. Athens, Greece, 2000.

[22] Hodges, Julia, Shiyun Yie, Ray Reighart and Lois Boggess (1996) "An automated system that assists in the generation of document indexes", *Natural Language Engineering* 2(2):137-160.

[23] Jacquemin, Christian, Judith L. Klavans and Evelyne Tzoukermann (1997) "Expansion of multi-word terms for indexing and retrieval using morphology and syntax", *Proceedings of the 35th Annual Meeting of the Assocation for Computational Linguistics*, (E)ACL'97, Barcelona, Spain, July 12, 1997.

[24] Justeson, John S. and Slava M. Katz (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering* 1(1):9-27.

[25] Klavans, Judith, Nina Wacholder and David K. Evans (2000) "Evaluation of Computational Linguistic Techniques for Identifying Significant Topics for Browsing Applications" *Proceedings of LREC*, Athens, Greece**.**

[26] Klavans, Judith and Philip Resnik (1996*) The Balancing Act*, MIT Press, Cambridge, Mass.

[27] Klavans, Judith, Martin Chodorow and Nina Wacholder (1990) "From dictionary to text via taxonomy*", Electronic Text Research*, University of Waterllo, Centre for the New OED and Text Research, Waterloo, Canada.

[28] Larkey, Leah S., Paul Ogilvie, M. Andrew Price, Brenden Tamilio (2000) Acrophile: An Automated Acronym Extractor and Server In Digital Libraries *'Proceedings of the Fifth ACM Conference on Digital Libraries*, pp. 205-214, San Antonio, TX, June, 2000.

[29] Lawrence, Steve, C. Lee Giles and Kurt Bollacker (1999) "Digital libraries and autonomous citation indexing*", IEEE Computer* 32(6):67-71.

[30] Milstead, Jessica L. (1994) "Needs for research in indexing", *Journal of the American Society for Information Science*.

[31] Mulvany, Nancy (1993) *Indexing Books*, University of Chicago Press, Chicago, IL.

[32] Nevill-Manning, Craig G., Ian H. Witten and Gordon W. Paynter (1997) "Browsing in digital libraries: a phrase based approach", *Proceedings of the DL97*, Association of Computing Machinery Digital Libraries Conference, 230-236.

[33] Paik, Woojin, Elizabeth D. Liddy, Edmund Yu, and Mary McKenna (1996) "Categorizing and standardizing proper names for efficient information retrieval",. In Boguraev and Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge, MA.

[34] Wall Street Journal (1988) Available from Penn Treebank, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[35] Tolle, Kristin M. and Hsinchun Chen (2000) "Comparing noun phrasing techniques for use with medical digital library tools", *Journal of the American Society of Information Science* 51(4):352-370.

[36] Voutilainen, Atro (1993) "Noun phrasetool, a detector of English noun phrases", *Proceedings of Workshop on Very Large Corpora*, Association for Computational Linguistics, June 22, 1993.

[37] Wacholder, Nina, Yael Ravin and Misook Choi (1997) "Disambiguating proper names in text", *Proceedings of the Applied Natural Language Processing Conference* , March, 1997.

[38] Wacholder, Nina (1998) "Simplex noun phrases clustered by head: a method for identifying significant topics in a document", *Proceedings of Workshop on the Computational Treatment of Nominals*, edited by Federica Busa, Inderjeet Mani and Patrick Saint-Dizier, pp.70-79. COLING-ACL, October 16, 1998, Montreal.

[39] Wacholder, Nina, Yael Ravin and Misook Choi (1997) "Disambiguation of proper names in text", *Proceedings of the ANLP*, ACL, Washington, DC., pp. 202-208.

[40] Wacholder, Nina, David Kirk Evans, Judith L. Klavans (2000) "Evaluation of automatically identified index terms for browsing electronic documents", *Proceedings of the Applied Natural Language Processing and North American Chapter of the Association for Computational Linguistics (ANLP-NAACL) 2000*. Seattle, Washington, pp. 302-307.

[41] Wright, Lawrence W., Holly K. Grossetta Nardini, Alan Aronson and Thomas C. Rindflesch (1999) "Hierarchical concept indexing of full-text documents in the Unified

Medical Language System Information Sources Map". *Proceedings of AMIA 1999*, American Medical Informatics Association, November, 1999.

[42] Yarowsky, David (1993) "One sense per collocation", *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, pp 266-271.

[43] Zhou, Joe (1999) "Phrasal terms in real-world applications". In *Natural Language Information Retrieval*, edited by Tomek Strazalowski, Kluwer Academic Publishers, Boston, pp.215-259.

Figure 1 Intell-Index opening screen < http://www.cs.columbia.edu/~nina/IntellIndex/indexer.cgi>

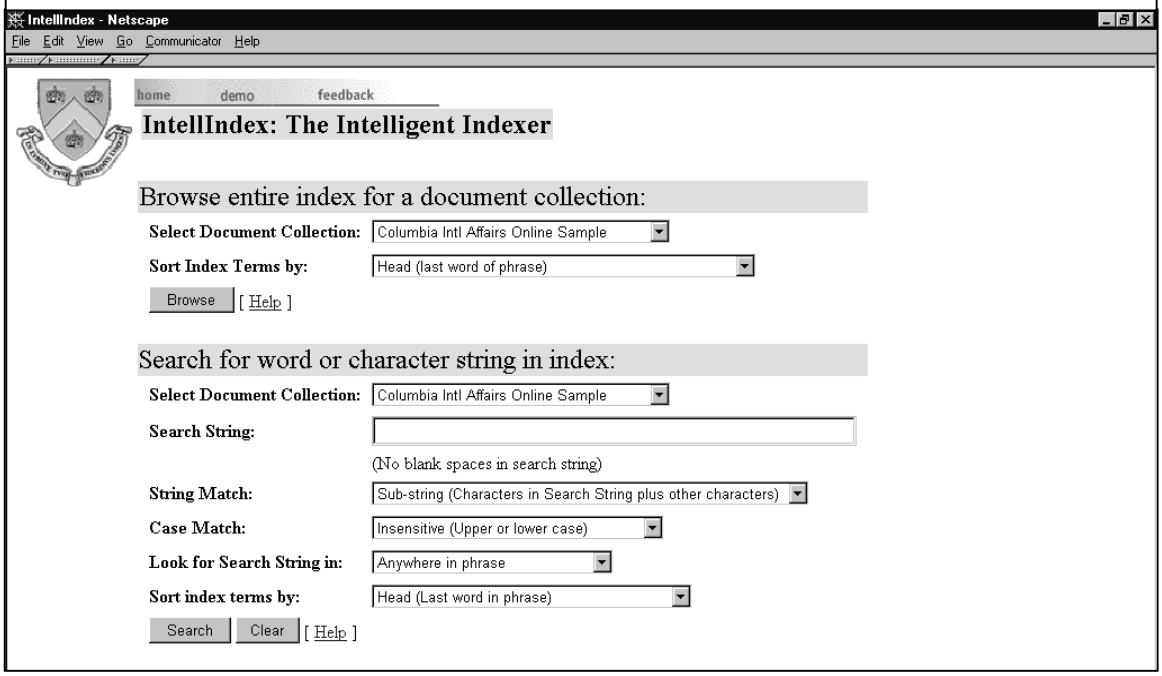

Figure 2 Browse term results

Browse Term Results

6675 terms match your query

ability
political ability

ABM

abuses human rights abuses

acceptance
widespread acceptance
broad acceptance

access full access
U.S. access

accession quick accession

accessions earlier accessions

accommodation
political accommodation

accomplishment
significant accomplishment

Accord Trilateral Accord

Accords Background De-Nuclearization Accords

accord
subsequent post-Soviet accord
bilateral accord
bilateral nuclear cooperation accord

accords nuclear-weapon-free-zone accords

accounting ...