Characteristics of Fulltext Index by Human and Automatic Indexing Systems

전문색인에 있어서 수작업 색인과 자동색인의 특성

Giyeong Kim*

ABSTRACT

The purpose of this study is to investigate the characteristics of indexes by human and machine, and differences between them in terms of term identification in a fulltext environment. A back-of-book index and two indexes produced by two term identifiers (LinkIt and Termer) as pseudo-indexing systems for a whole body of a monograph are examined. In the investigation, the traditional contrast between manual and automatic indexing is confirmed in fulltext environment; manual index is for browsing and human use, and automatic index is for searching and machine use. The border between them, however, becomes vague. Some considerations for the use of the term identifiers for browsing and for searching are discussed, and further research for the use of the term identifier is suggested.

초 록

본 연구는 전문(fulltext) 환경에서 수작업 색인과 자동색인의 색인용어의 특성과 차이점을 알아보는 것을 그 목적으로 한다. 이를 위해 영어로 작성된 단행본에 대한 권말색인과 두 개의 유사 색인 시스템(LinkIt 과 Termer)을 이용한 색인들이 이용되었다. 이러한 비교분석을 통해 수작업 색인은 이용과 브라우징에 대한 강점이 있으며 자동색인은 자동 시스템에서의 탐색에 강점이 있음을 확인하였지만, 양자간의 경계가 불분명해짐도 아울러 확인하였다. 마지막으로 브라우징과 탐색을 위한 유사 색인 시스템의 이용에 있어서 고려할 점과 이에 대한 향후 연구에 대하여 토의하였다.

Keywords: index term characteristics, manual index, automatic index, fulltext index
색인어 특성, 수작업 색인, 자동색인

---

* Assistant Professor, Dept. of Library and Information Science, Yonsei University (gkim@yonsei.ac.kr)

- Received : 24 May 2008 ▪ Revised : 30 May 2008 ▪ Accepted : 15 June 2008
- Journal of the Korean Society for Information Management, 25(2): 199-221, 2008. [DOI:10.3743/KOSIM. 2008.25.2.199]

1. Introduction

The main purpose of indexing is to construct representations of published items in a form suitable for inclusion in some types of database. This database of representation could be in printed form, in electronic form, or in card form (Lancaster 1998). Manual indexing has been known to be totally different from automatic indexing in their indexing process. Usually, manual indexing is an assignment of index terms from controlled vocabulary on the indexed item and automatic indexing is an extraction of terms from the text (Lancaster 1998). One of the differences between them is that terms from manual indexing are usually multi-word terms and abstracted, while terms from automatic indexing are single-word terms and extracted from texts directly. The answer for the question, which is better, has not been clear (Lancaster 1998; Hmeidi et al. 1997). The answer has been varied by the criteria. For instance, automatic indexing has been usually known to be more efficient than manual indexing, while manual indexing has been evaluated as good for index term quality. Also, Fidel (1994) pointed out that automatic indexing is document oriented rather than user-centered.

In the meantime, most of indexes in the research have been made from certain types of substitutes of documents, such as titles and abstracts. It means that the ways of and the characteristics of the indexing would not appropriate for other type of text, such as fulltext. As the contents in the Internet is growing exponentially, the need of fulltext indexing is also growing, because there is a number of fulltexts in electronic forms which is to be searched. So far, the indexing in the Internet has mainly not based on fulltext but other features in the Internet documents. Yang (2005) points out that the ad-hoc methods in Text Retrieval Conference (TREC), which is commonly based on the traditional automatic indexing, might not appropriate for the web retrieval. He identified that the most common approaches in the web retrieval are the link-based approach, citation-based approach, and organizational approach, which is mainly carried out with document categorization. That is, the current web retrieval is mainly based on the substitute indexing.

There has been criticism both on automatic indexing and on manual indexing. Yang (2005) also emphasized certain weaknesses in retrieval approaches based on single sources of evidence in the same review. This means that there is a plenty of room for the development of the web retrieval, and possibly for our attention to fulltext. East (2005) finds out that the indexing in the database is inadequate for monographs, and stresses on the need of improvement by automatic generation of subject-rich document surrogates or by author-generated metadata while investigating an index database for monographs. That is, the current index databases with small number of terms assigned to an item are not appropriate to monographs, and they need additional features for representation of the indexed item.

Some research questions arise based on the dis-

cussion above: what are the characteristics of manual and automatic indexing in fulltext environment? How much are they different from each other? And which is better than the other? In order to find out the answers, we will analyze a back-of-book (BOB) index, because it could be a case of manual fulltext indexing, and because we can get indexes by automatic fulltext indexing if we have a fulltext document in electronic form. On the other hand, the development in the domain of Information Retrieval and Natural Language Processing makes identification of multi-word terms from texts available. We attempt to identify some characteristics of fulltext indexes by manual and by automatic indexing systems. Two indexing systems are selected, LinkIT and Termer. Even though they are not developed as indexing systems but noun phrase (NP) chunker or term identifier, the terms they identify are more similar to the ones identified by human than by traditional automatic indexing systems. Thus, they would be called pseudo-indexing systems. While comparing indexes by human and by the systems, some important characteristics are identified.

There are some criteria in evaluation of index. The criteria can be divided into 3 different categories: retrieval system evaluation, usability testing, and value of indexing (Milstead 1994). In retrieval system evaluation, quality of index is evaluated through the results from the evaluation of a retrieval system, such as recall and precision. It is measurable but indirect way. That is, there would be other factors than the index, which affect the retrieval system evaluation. Moreover, this has been mainly based not on fulltext but on substitutes of texts. In usability testing, the quality of an index is evaluated by end users in certain criteria, such as efficiency and effectiveness. It is useful to identify index quality, but gives little information about the relationship between the results from the testing and the characteristics of indexing. That is, we cannot know which characteristics affect the results of the testing. In value of indexing, the index is described with some features. This shows characteristics of an index, but cannot provide the decision which is better or worse. In this study, the third is adopted, because the purpose of this study is not on putting the methods of indexing in an order, because this can show the characteristics of the indexes thoroughly, and because we don't have an agreement to the criteria for the goodness of index yet.

In this paper, we focus on describing characteristics of manual and automatic indexing for a fulltext document. In the next section, we discuss the characteristics and evaluation of BOB. In the methodology section, we describe the characteristics of selected indexes, and the way to compare the indexes with each other in this study. Then we report some statistical characteristics from the comparison and present discussions on the results. Finally, a summary of this investigative study is provided.

2. Theoretical Background

2.1 Differences between Manual and Automatic Indexing

For the differences between manual and automatic indexing, Anderson and Pérez-Carballo (2001a, b) provide a good review. They summarize the differences in 8 features of indexing as follows.

1) Size of documentary unit: Manual indexing tends to focus on larger documentary unit.
2) Extent of indexable matter: Automatic indexing is now routinely based on the complete text, whereas much manual indexing may be limited to an abstract or other summarization of the complete text.
3) Exhaustivity: Automatic indexing tends to be exhaustive, considering most words in indexable matter as potential indicators of terms.
4) Specificity: Automatic indexing tends to use very specific terminology, because it uses the actual language of the text.
5) Browsable displayed index: Browsable displayed indexes with multi-term context-providing headings are certainly possible with automatic indexing, but they are not as common as with manual indexing.
6) Searching syntax, display syntax: For the automatic indexing, there are so many sophisticated techniques for selecting, combining, manipulating and weighting terms, On the other hand, the syntactic possibilities for the combination of terms to create context providing headings are richer for manual indexing.
7) Vocabulary management: manual indexing is stronger in vocabulary management using cross references linking synonymous terms, pointing to related terms, and distinguishing among ambiguous homographs.
8) Surrogation: For manual indexing, this is connected to the amount and style of information provided in index headings. For automatic indexing, this is related to the size and style of the documentary unit records.

Most of these differences are not from their inherent features but from the tendency of their products. Some techniques for both indexing methods would affect the differences, such as the use of controlled vocabulary in automatic indexing and indexing policy, which focuses on exhaustivity and/or specificity, in manual indexing. In addition, Rasmussen (1994) also characterizes the differences between the relative effectiveness of controlled vocabulary and free text.

2.2 Back-Of-Book Index

For BOB index, the features for the value of indexing can be divided into 4 categories in evaluating BOB index: introductory note; physical format, typology and style; content of the index; and structure and accuracy of the index entries (The American Society of Indexers 2007). The introductory note

should be provided if any aspect of the index requires explanation. The format, typography and style are for providing maximum ease of scanning the index and locating individual entries. The second category provides us that the purposes of the BOB index are helping readers to scan or to browse the index terms and to locate individual terms. For the third category, content, all significant items in the text must appear in the index. The index must bring together references to similar concepts that are scattered in the text, or that are expressed in varying terminology. This can be done by establishing a single heading with subheadings, by using cross-references, or with other devices. This means that identifying significant items is most important, and the relationships between the items, such as parent-child (heading - subheading) relation, and association ("see" and "see also"), should also be represented in the indexing. For the last category, structure and accuracy, the index entries should be arranged in a recognizable order, such as alphabetical, classified, chronological, or numeric order. This would be closely related to the second category, because the structure and accuracy is mainly for ease of locating entries.

The four categories are the criteria in evaluating BOB indexes for the ASI / H. W. Wilson Award, which honors excellence in indexing of an English language monograph or other non-serial publication (The American Society of Indexers 2007). In evaluation of the BOB indexes for the award, five human judges select a winner with the four criteria. Among them, the introductory is the own feature of BOB index than other indexes, and the second (Physical format, typology and style) and the last categories (structure and accuracy) are about editing index in a physical form. The only criterion for the content of the index is the identifying index terms, which could be used for comparing human BOB index with other types of indexes.

Gratch and his colleagues (1978) investigate the characteristics of BOB indexes in 113 books in nine disciplines in the humanities and social sciences. The identified characteristics are length statistics, arrangement, scope, headings, subheadings, locators, control devices, and physical appearance. Among them, the length statistics is the only one related to term identification. Most of them are related to physical forms of the index. In addition, they result that the more number of pages in a book, the less index density, which is the average number of terms per a page in a book. This means that manual indexers tend to fit the number of index terms to a certain number. Finally, they recommend that pre-coordinated terms should be used instead of single words for index entries. That is, multi-word terms are better than single word terms for index entries.

2.3 Evaluation of BOB Index

American Society of Indexers suggests a checklist for BOB index evaluation (American Society of Indexers 2006). The list constitutes of eight criteria: reader appropriateness, main headings, subheadings, double postings, locators, cross-refer-

ences, length and type, and format. Reader appropriateness is that the indexed terms should appropriate for the intended audience. Main headings should be relevant to the needs of the reader, be pertinent and specific, and have not more than 5-7 locators. Subheadings should be useful, concise, and coherent. The number of subheadings is about right. Subheadings should be double posted. The locator should be accurate. Cross-references, such as *see* and *see also,* should be provided. The length of index is 3-5% of the content indexed. These criteria can be categorized into 3 groups: term identification (reader appropriateness, main headings, and subheadings), relationship assignment (subheadings, double postings, and cross-reference), and physical features (locaters, length and type, and format).

There are some attempts to investigate the features of good BOB indexes. Wittman (1990) compares good BOB indexes, which had won the Wheatley Medal, with bad indexes using quantitative methods. The unit of analysis is a subheading of the indexes. Previous comparison suggested that subheadings in award-winning indexes are vivid and concise, conveying contents of a indexed text in a few words which are the essence of the text; others were often cryptic, rambling, and vague. Then the author questions whether this subjective impression of qualitative differences could be formulated on the basis of such objective and quantifiable features as (1) subheading length, (2) initial word of subheadings, (3) syntactic relationship between subheadings and main headings, and (4) semantic relationships between subheadings and the text indexed. Each criterion is measured as follows.

1) Subheading length: Number of words in a subheading
2) Initial word of subheading: Part-of-speech of the first word in a subheading
3) Syntactic relationship between subheadings and their main headings: Categorized as direct, indirect, topical by the researcher
4) Semantic relationships between subheadings and the text indexed: Categorized as exact match, near match, synonymous match, and paraphrase by the researcher

As a result, the subheading length of good BOB indexes is 2.3 and it is not so different from the bad indexes (2.6). The initial word of subheadings is typically a noun, a verb, or a preposition (87%). In the syntactic relationship between subheadings and main headings, topical category is major for good indexes and this is a big difference from the bad indexes (18%). In semantic relationships between subheadings and the text indexed, exact or near match categories are major (55%) for good indexes and this is a little bit different from bad indexes (36%).

While investigating the use of automatically identified phrases as index terms in a dynamic text browser, Wacholder and her colleagues (2001) suggest 3 criteria for evaluation of index terms: coherence, thoroughness of coverage of document content, and usefulness. Coherence is measured

as number of junk terms that humans readily recognize as incoherent, and is related to the term identification performance of automatic indexing systems. Thoroughness of coverage of document content is measured as a ratio between number of index (identified) terms and the size of the text indexed, and is related to indexing exhaustivity. Usefulness is measured as users' perceptions of the usefulness of index terms, and is related to usability test. The three criteria focus on evaluating automatic index.

While investigating exhaustivity, which defined as the number of terms assigned to a document, Sparck-Jones (1973) suggests that there is an optimal level of exhaustivity for a particular collection. It means an index with high exhaustivity would not be a good index. The indexing exhaustivity is important, but closely related to the request exhaustivity and term specificity in information retrieval system. This is the reason that it is difficult to show the only effect of index exhaustivity on a search performance.

Lathrop and his colleagues (1997) identify 5 criteria for usable index in documentations for software products; orientation, retrievability, analysis, flexibility and learnability. Orientation means that an index should orient its readers in terms of the terminology and concepts covered in the document. Retrievability is an ability to make users find the information they need easily. Analysis means that an index should provide a good topic analysis of the text. Flexibility is that an index provides adequate entries for its readers. And learnability means that an index helps educate the readers. These five criteria enhance the effectiveness, efficiency and enjoyment in terms of the products usability.

Hert et al. (2000) suggest some measures of performance for online indexing structures in the networked environment. The measures broadly divided into 4 categories; usability, efficiency, effectiveness, and satisfaction. The satisfaction is measured by user's post-questionnaire. The efficiency is estimated with 4 sub-measures: 1) Average number of user clicks per task compared across index conditions, 2) Average number of searches initiated per task compared across index conditions, 3) Average time spent in the index for each task compared across index conditions, and 4) Average time spent in the index page per visit compared across index conditions. The effectiveness also consists of 2 sub-measures: average success rate for each task determined by the researchers and compared across index conditions and average user perception of success for each task compared across index conditions. The usability is a meta-category, and estimated from the 6 sub-measures of the efficiency and the effectiveness. <Table 1> lists the criteria for index evaluation discussed above.

So far, we review a variety of criteria in evaluation of indexes in various indexing environments. The criteria are bases for investigation of the characteristics in manual and automatic indexes in the study.

〈Table 1〉 Criteria for Index Evaluation

Name	Measured Feature	Scale	Availability*	Developer
Subheading Length	# of words in a subheading	ratio	both	Wittman (1990)
Initial word of subheadings	POS of initial word in subject heading	3 categories : n/v/prep, adj/adv, conj/art/prep	both	Wittman (1990)
Syntactic relationship	relationship between subheadings and main headings	3 categories: direct, indirect, topical	manual index	Wittman (1990)
Semantic relationship	relationship between subheadings and the text indexed	4 categories: exact match, near match, synonymous match, paraphrase	both	Wittman (1990)
Orientation	Does an index orient its readers to the terminology and concepts covered in the documentation?	Dichotomous/Ratio	both	Lathrop, et. al. (1997)
Retrievability	How long does it take users to find the information they need?	Ratio (Time)	both	Lathrop, et. al. (1997)
Analysis	Does the index provide a good topic analysis of the text?	Dichotomous/Ratio	Both	Lathrop, et. al. (1997)
Flexibility**	Does the index provide adequate entries for readers with varying levels of expertise to access the information they need?	Dichotomous/Ratio	both	Lathrop, et. al. (1997)
Learnability	Does the index help educate readers? Does it shorten the time required for them to be more productive? Does it help readers understand relationships of concepts, tasks, and procedures?	Dichotomous/Ratio	both	Lathrop, et. al. (1997)
Indexing Exhaustivity	number of index terms, topics, and/or themes	Ratio	both	Spark-Jones (1973)
Term Specificity	Average number of posting of terms in a document	Ratio	both	
Coherence	number of junk terms that humans readily recognize as incoherent	Ratio	both	Wacholder, et. al. (2001)
Thoroughness of coverage of document content	ratio between number of index (identified) terms and the size of original text indexed	Ratio	both	Wacholder, et. al. (2001)
Usefulness	user's perceptions of the usefulness of index terms.	Ordinal/Ratio	both	Wacholder, et. al. (2001)
Efficiency1***	Average number of user clicks per task compared across index conditions	Ratio	both	Hert, et. al. (2000)
Efficiency2***	Average number of searches initiated per task compared across index conditions	Ratio	both	Hert, et. al. (2000)
Efficiency3***	Average time spent in an index for each task compared across index conditions	Ratio	both	Hert, et. al. (2000)
Efficiency4***	Average time spent in an index page per visit compared across index conditions	Ratio	both	Hert, et. al. (2000)
Effectiveness1***	Average success rate for each task determined by the researchers and compared across index conditions	Ratio	both	Hert, et. al. (2000)
Effectiveness2***	Average user perception of success for each task compared across index conditions	Ratio	both	Hert, et. al. (2000)

* Criterion's availability for manual index or automatic index
** same to Reader Appropriateness by ASI
*** the six criteria makes a meta-category, usability, and the six are for networked information

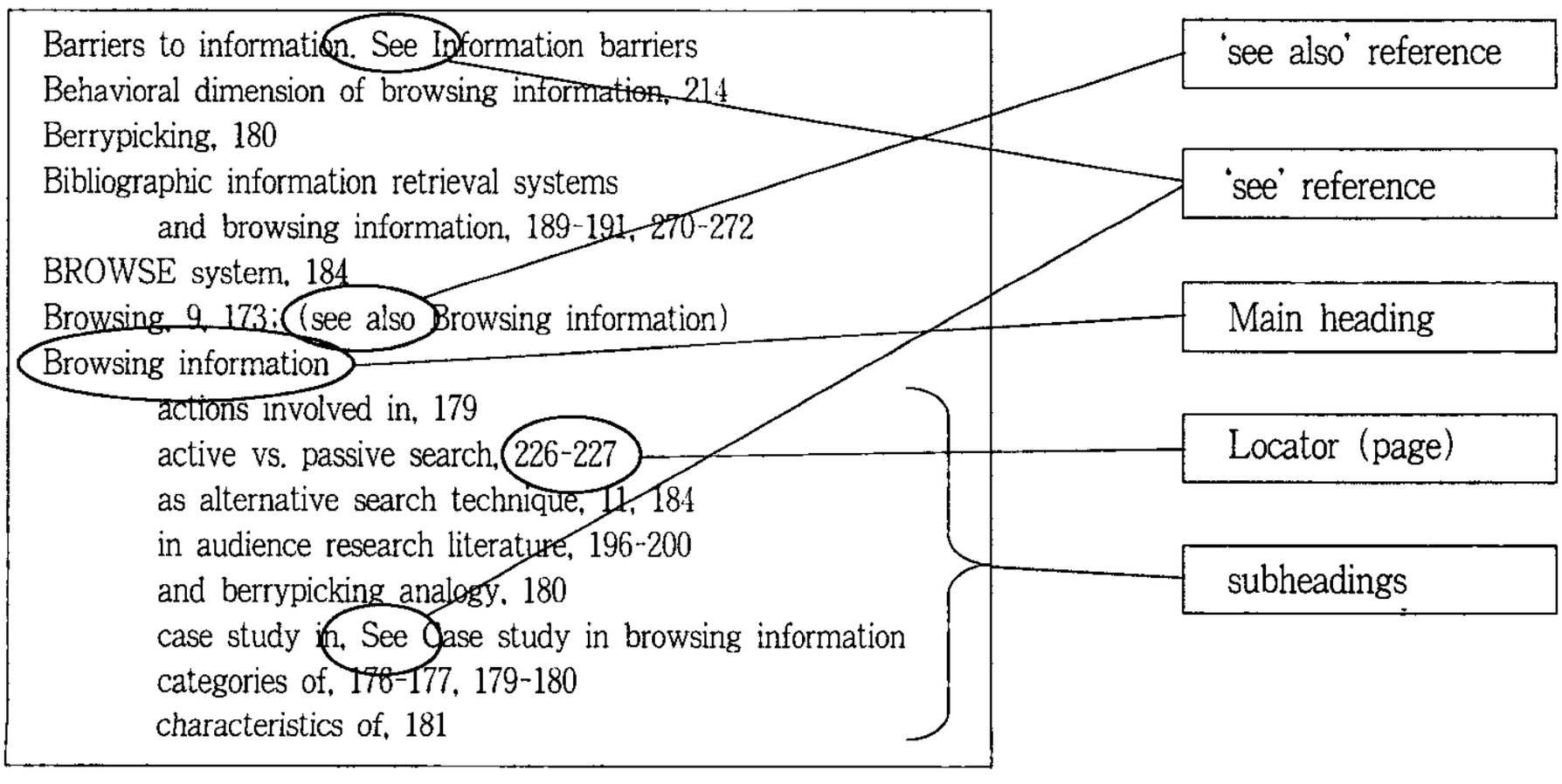
3. Methodology

In this study, 3 indexes for a fulltext document are characterized and compared with each other. The name of the book is "Accessing and Browsing Information and Communication: An Interdisciplinary Approach" by Rice, R.E.; M. McCreadie; and S. Chang (2001). As titled, its content is a study in a domain of Social Science. The reason that it is chosen is it has well-organized BOB index, and the authors provide it in electronic formats (*ASCII* and *rtf*) as data for this study. It has a title and bibliography, a preface, a table of contents, fifteen chapters, references, and a BOB index. The fifteen chapters are the object of the BOB index, because the page number in the index starts from 1, and the page 1 is the first page of chapter 1. The fifteen chapters have 106,788 words, excluding tables and figures.

The focus of this study is the identification of terms as indicators of the content in the fulltext document. In the BOB index, there are other features then terms; levels (main and subheading), links between terms ('see' and 'see also'), and various relationships between main and subheading. Also, the automatic indexes provide other information than terms. To compare them with each other in terms of term identification, the other features are removed. The following is the characteristics of and manipulation in each index.

3.1 Original BOB Index by Human

The BOB index was made by an amateur indexer, who was a student in a program of Master of Library Science in U.S. After the indexing, the authors examined the indexed terms. In the final product of BOB index, there are 1,066 identified terms (main and subheadings) with 2,785 words. The number of words per term is 2.61, and the range



〈Figure 1〉 A Part of BOB index

of the number is 1 (for example, *Assessment*) to 8 (for example, *Resources in framework for theory of browsing information*). The index is constructed with 2 levels: main heading and subheading. Among the terms, 323 terms are main headings and others (743) are subheadings. Not all main headings have subheadings; only 88 main headings have their subheadings (8.44 subheadings per main heading). It also has linking devices; 58 *see* and 16 *see also*. The indexing density, which is an average number of index terms per pages (Gratch et al. 1978), is 3.42. As usual, the locator is the page number.

In <Figure 1>, every feature of the BOB index is shown: main headings with or without subheadings, locators, levels, *see* and *see also* references. 'See' reference represents synonymous and/or equivalent relationships, and 'see also' reference does association. The relationship between main heading and its subheading seems not to be clear, and in subheadings there are devices to represent the relationships. For instance, a subheading in <Figure 1> "*as alternative search technique*" is constituted with a term (*alternative search technique*) and a device (*as*), which shows the relationship between the subheading and its main heading. The devices are usually attached in front of or in rear of the terms, and are usually prepositions, but not all. These devices in the subheadings are called 'linking device' in this article, and are removed to identify terms. The identified linking devices are: *and, as, between, by, for, from, in, involved in, of, on, regarding, to, and with.* The linking devices are removed in order to get a list of terms. After removing, the number of identified index terms are 706, and 1,726 words are used (2.44 words per term).

The locator, page number, is not fixed in an electronic form of the document, so that it is not a good locator for comparing with other types of index. There would be other locators, which are fixed in any form, such as a whole book, a chapter, a section, a paragraph, and a sentence. The locater should detail enough so that a reader can find the term easily. But the terms in the BOB index has only page information as the locator. Moreover, most of terms by human are not extracted but abstracted. That means we cannot know the exact location of a term with more detailed unit of locater than page. Therefore, we cannot change the locator from page to paragraph or sentence. For the section, the definition of it is not logically consistent in the book. That means some chapters have many sections, but others are not. In addition, the section does not correspond to the page. That means a section may starts and ends in the middle of a page. Thus, the page cannot be changed to the section. Finally, the only candidate is the chapter. The chapters always begin with new pages, so it corresponds to the page. The book has 15 chapters, and there are 20.8 pages and 7,119.2 words per chapter. For the comparison between indexes, the book is divided into 15 chapters, and the terms are re-assigned to the chapter numbers according to their page numbers. For example, '*berrypicking (page 180)*' in <Figure 1> is assigned to Chapter

9, because Chapter 9 is from page 169 to 216.

3.2 The Index by LinkIt System

The other two indexes are produced by automatic indexing systems: LinkIt and Termer. LinkIt is a noun phrase (NP) identifier developed by Klavans, J. and Wacholder, N. in Columbia University (Wacholder 1998, Evans 1998). The basic role of LinkIt is to identify simplex NPs as significant topics within a document from pre-processed text by a part-of-speech (POS) tagger. Simplex NP is nominal elements, which together constitute a simplified representation of the content of a document. A simplex NP is a maximal NP that includes premodifiers such as determiners and possessives but not post-nominal constituent such as prepositions or relativizers. An important property of these simplex NPs is that the phrasal head is the last element. Some examples of simplex NP are *political participation and technological choices*. Simplex NP can be contrasted with complex NP, such as *political participation with implications for democracy* where the head of the NP is followed by a preposition, or *technological choices related to preservation* where the head is followed by a participle verb.

The input to LinkIt is a pre-tagged text by the Alembic Workbench by the MITRE group (Evans 1998). It performs POS tagging and also identifies named entity for people, organizations, and locations. The existence of the pre-processing system would affect the performance of LinkIt. The version 4.40 of it and LinkIt are installed on Sun OS 5.8. <Figure 2> shows a part of the output from LinkIt.

LinkIt creates 3 files, ending with (*filename*).np, (*filename*).stat, and (*filename*).stat2. The np file is the standard output file. This file contains a list of the NPs from the document, and also a clustered list of those NPs. The stat and stat2 files contain statistical characteristics, such as the frequencies of the part of speech tags in different formats. The output in <Figure 2> is np file. In the np file NPs are listed with information of the sentence it occurred in, the token span, and probable information about relationships to other NPs. It includes the information on whether the NP is in apposition, is a possible head or possible modifier of another NP, and previous occurrences of words in the NP. The format is as follows:

S41 801-803 (232) the public (pocc: 128.443) library (pocc: 230.796) [1]

- S# is the sentence number the NP occurs in.
- The next two digits refer to the token span of the noun phrase, where [the] is the 801st, [public] is the 802nd, and [library] is the 803rd token.
- The number in parentheses is a unique identifier assigned to this simplex NP.
- The number in [ ] is the term frequency in the text.

```
File ../../tools/New-Intell-index/db/ricetext__LinkIT/Chapter11.tagged processed on Wed
Apr 24 15:41:50 2002


------------------------------------------------
II: Noun Phrases Ordered by Heads:
------------------------------------------------

LIBRARIES
        S4 51-51 (14) libraries [8]
        S41 794-796 (230) the academic library (pocc: 229.792) [1]
        S41 801-803 (232) the public (pocc: 128.443) library (pocc: 230.796) [1]
        S41 809-811 (234) the special library (pocc: 232.803) [2]
.....

------------------------------------------------
III: Words as heads and mods:
------------------------------------------------


library:
 head:
        S4 51-51 (14) libraries
        S22 471-471 (135) libraries (pocc: 14.51)
        S37 746-747 (215) across libraries (pocc: 135.471)
        S41 792-792 (229) libraries (pocc: 215.747)
        S41 794-796 (230) the academic library (pocc: 229.792)
        S41 801-803 (232) the public (pocc: 128.443) library (pocc: 230.796)
        S41 809-811 (234) the special library (pocc: 232.803)
        S42 820-822 (236) the academic (pocc: 230.795) libraries (pocc: 234.811)
.....

_____Timing Information
Took 290000 ticks (0.290000 seconds)
finish: 290000 start: 0
```

〈Figure 2〉 A Part of the Output from LinkIt

If a noun phrase is possibly in apposition with another noun phrase, that will be marked with a (papp: #) tag. Similarly, (phead: #) and (pmod: #) specify that this noun phrase is a possible head or modifier of another noun phrase. In all cases the # refers to the unique number for the NP given in the first parenthesis. For each of the words of the noun phrase, there might be a (pocc: A.B) label. It denotes a previous occurrence of the word in noun phrase A, token number B

As shown in <Figure 2>, the output divided into 2 parts, “noun phrases ordered by heads” and

"words as heads and mods", except for the transaction information on the top of the file, and timing information on the bottom of the file. An identified term list by the system will be extracted from one of the parts, NP list ordered by heads. First, the other parts than the noun phrases are removed. Then, the heads are removed. After that, other information than the terms (S#, two digits in the parenthesis, the unique identifiers, and the term frequencies) are removed. Because of incompleteness of the POS tagger, the remained terms would include inappropriate words, such as 'of' in 'of Information'. The inappropriate words are usually located in front and rear of the terms. 49 words are identified as the inappropriate (see Appendix 1), and removed. Finally, an index term list in alphabetical order is produced.

3.3 The Index by Termer System

Termer is an experimental term identification program and coded originally by Katz, S.M. based on an algorithm for identifying technical terminology (Justeson & Katz 1994). The common grammatical property of the technical terms from dictionaries of technical vocabulary is identified as multi-word NP, which mainly consist of nouns, adjectives, and preposition 'of'. But all NPs are not terminological NP; it differs from other NPs because they are lexical. Lexical NPs are subject to a more restricted range and extent of modifier variation, on repeated references to the entities they designate, than are non-lexical NPs. Based on the observation, an algorithm for identifying technical terms from texts is developed: Candidate strings must occur twice or more times in the text, and are those multi-word NPs, which are consist of adjectives (not determiners), lexical nouns (not pronouns), and prepositions. Some typical examples are as follows, where A is an adjective, N is a noun, and P is a preposition.

- AN: popular art, informal communication
- NN: information communication, tv commercial
- AAN: complex electronic environment, experimental interactive system
- ANN: personal interest profile, visual communication process
- NAN: databases similar findings, users specific reference
- NNN: information science field, subject behavior information
- NPN: control of culture, goals of users

The program in this research is revised version 1.0, coded by Song, P. from version 3.6 by Min-Yen Kan. The difference between them is in the input file; input for original Termer is tagged text by Comlex, whereas one for revised Termer is tagged text by Alembic Workbench 2.8, which is the same preprocessor for LinkIt. The program is written in Perl, and installed on Sun OS 5.8. <Figure 3> is a part of the output file.

```
Filename /raid/users/nlp/tools/New-Intell-index/db/ricetext__Termer/Chapter1.tagged
processed on Tue Apr 16 21:24:28 EDT 2002

-------------------------------------------------
II: Noun Phrases Ordered by Heads:
-------------------------------------------------

able
        S9 26-27 (33) not able

access
        S26 17-18 (110) understanding access
        S42 25-26 (162) partial access
        S52 12-14 (184) conscious decisions access
        S54 36-37 (188) privileged access
        S54 106-107 (195) when access
        S55 17-18 (198) privileged access
        S60 1-3 (217) when others access
        S65 2-3 (235) struggle access
.....
```

〈Figure 3〉 A Part of the Output from Termer

The output file can be divided into 2 parts; transaction information on the top of the file and the list of NP ordered by heads. First, the transaction information and the title of the list are removed. Next, the heads only items are removed, because the focus of the research is on the identified technical terms. Each item of the list is as follows.

```
S158 16-18 (443) explicit browsing capability
```

- S# is the sentence number the NP occurs in.
- The next two digits refer to the token span of the noun phrase in the sentence, where [explicit] is the 16^{th} token, [browsing] is the 17^{th}, and [capability] is the 18^{th} token in the 158^{th} sentence.
- The number in parentheses is a unique identifier assigned to this technical term.

To list terms only, the other information is removed, and because of the same reason to LinkIt, the inappropriate words are removed. Then the terms are ordered alphabetically.

3.4 Comparisons between the Indexes

After extracting terms from three indexes, we have three term lists for each chapter: (1) a list of terms identified as representatives of the text by human (BOBL); (2) a list of terms identified as NP by LinkIt (LINKL); and (3) a list of terms

identified as technical terms by Termer (TERML).

Direct comparisons between the term lists are difficult, because the terms in all the lists are basically multi-word terms, and because the terms by human are not only extracted but also abstracted. <Table 2> shows parts of the three lists for chapter 1.

In comparison among them, we cannot employ exact matching method. The other way to compare is partial matching method. For example, '*access*' in LINKL is matched to '*outsiders and access to information*' in BOBL, because the term in BOBL includes the term *access* (LINKL → BOBL). However, the inverse matching cannot be identified, because there is no term that includes the term *outsiders and access to information* in LINKL (BOBL → LINKL). Therefore, there are six matching processes: BOBL → LINKL, BOBL → TERML, LINKL → BOBL, LINKL → TERML, TERML → BOBL, and TERML → LINKL. The term extracting from the indexes and matching is done by small programs written in Perl and tcsh on Sun OS 5.8.

4. Results and Discussions

<Table 3> shows some statistical characteristics of the three indexes. In BOBL, 866 terms are identified from the 15 chapters. This number is, as expected, much smaller than the numbers in other indexes, such as 15,883 terms in LINKL and 12,890 terms in TERML. In terms of the number of words in a term, BOBL is 2.25, which is bigger than others: 1.58 in LINKL and 1.91 in TERML. Even though the number of BOBL is bigger than others, the differences between them are not so significant. Moreover, there are many numbers identified

〈Table 2〉 Parts of the Three Term Lists for Chapter 1

Human BOB index (BOBL)	Index by LinkIt (LINKL)	Index by Termer (TERML)
Accessing information	abstracts	academic disciplines
browsing information	academic disciplines	academic research tradition
care-taking	academic research tradition	access
communication	access	access framework
compared with browsing	access framework	access information
equality issues	access issues	access information access relevant interpretations
fundamental behavior	access points	access information systems
information searching	access-related issues	access issues
information technology	accessing	access issues user
information-seeking process	account	access many different areas
methodology	account various users	access necessarily
outsiders and access to information	accounting	access necessary
role in information-seeking process	acquisition	access other individuals communication media
searching	action	access pertinent information
.....	activities	.....
	.....	

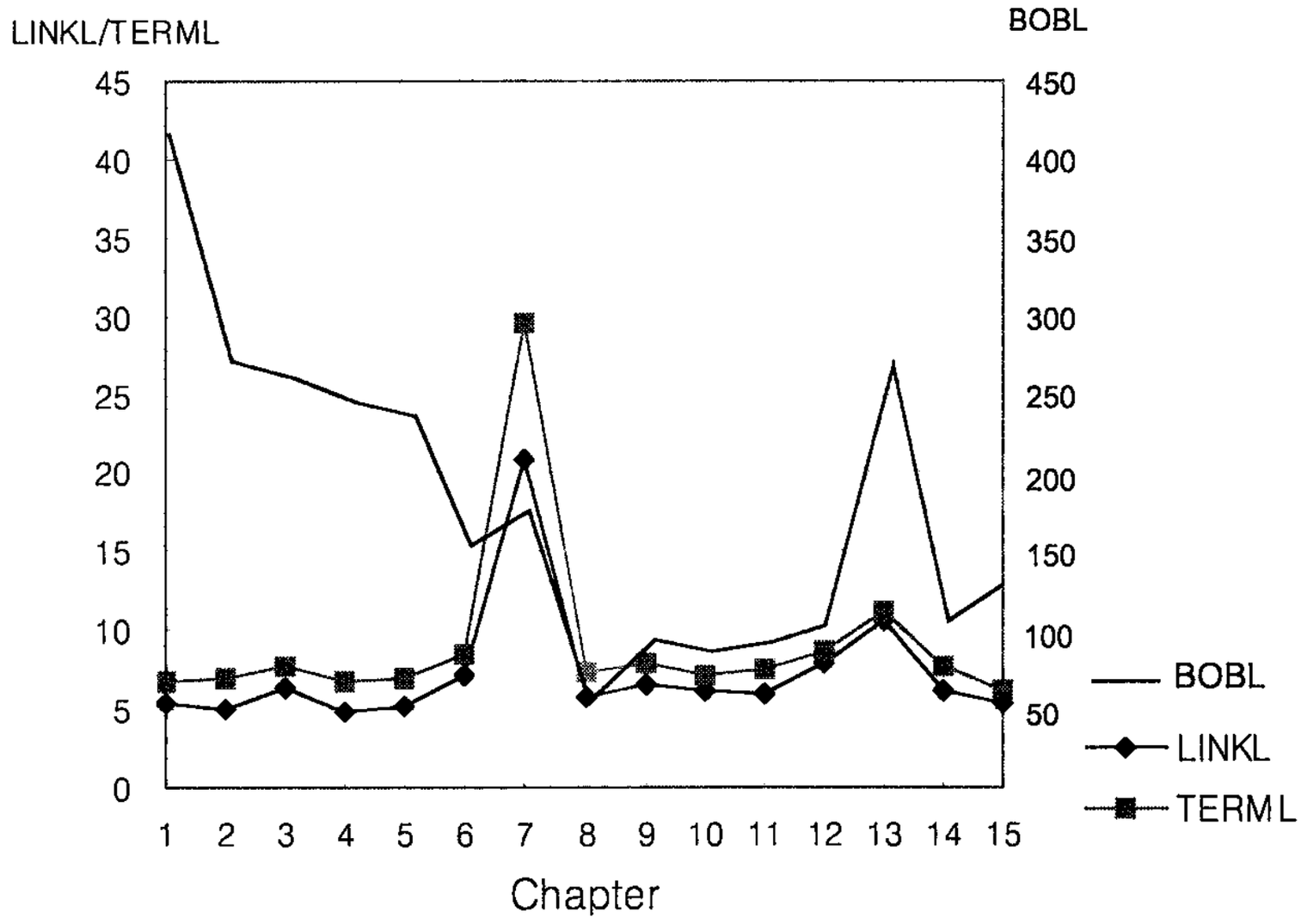
〈Table 3〉 Basic Statistics for the Fulltext and Three Indexes

	No. Words in fulltext	BOBL				LINKL				TERML			
		No. Terms	words in fulltext per term	No. Words	words per term	No. Terms	words in fulltext per term	No. Words	words per term	No. Terms	words in fulltext per term	No. Words	words per term
Ch 1	5860	14	418.57	34	2.43	1093	5.36	1677	1.53	877	6.68	1703	1.94
Ch 2	8285	31	267.26	75	2.42	1679	4.93	2597	1.55	1206	6.87	2311	1.92
Ch 3	14507	56	259.05	124	2.21	2306	6.29	3611	1.57	1888	7.68	3685	1.95
Ch 4	3162	13	243.23	31	2.38	658	4.81	957	1.45	465	6.80	875	1.88
Ch 5	3749	16	234.31	58	3.63	708	5.30	1078	1.52	537	6.98	1052	1.96
Ch 6	6343	42	151.02	125	2.98	897	7.07	1307	1.46	741	8.56	1351	1.82
Ch 7	10237	58	176.50	146	2.52	491	20.85	714	1.45	347	29.50	628	1.81
Ch 8	3866	76	50.87	188	2.47	678	5.70	1043	1.54	525	7.36	1049	2.00
Ch 9	15756	220	71.62	438	1.99	2406	6.55	4256	1.77	1976	7.97	3845	1.95
Ch 10	7674	127	60.43	260	2.05	1245	6.16	2010	1.61	1071	7.17	2023	1.89
Ch 11	3681	46	80.02	109	2.37	619	5.95	964	1.56	491	7.50	911	1.86
Ch 12	5012	54	92.81	116	2.15	636	7.88	971	1.53	573	8.75	1077	1.88
Ch 13	9953	37	269.00	93	2.51	940	10.59	1508	1.60	885	11.25	1704	1.93
Ch 14	2950	28	105.36	57	2.04	475	6.21	702	1.48	383	7.70	670	1.75
Ch 15	5753	48	119.85	96	2.00	1052	5.47	1670	1.59	925	6.22	1745	1.89
Sum	106788	866		1950		15883		25065		12890		24629	
Average	7119.20	57.73	123.31	130.00	2.25	1058.87	6.72	1671.00	1.58	859.33	8.28	1641.93	1.91

as terms in LINKL (805 times). If these are removed, the gap between the numbers of LINKL and BOBL becomes smaller. This means that LinkIt and Termer as automatic indexing systems can identify terms like a human indexer, at least in terms of the term length. But still there is a big difference between manual index (BOBL) and automatic indexes (LINKL and TERML). <Figure 4> shows average number of words in fulltext per term. This number is similar to the indexing density (Gratch et al. 1978), which is average number of index terms per pages.

The meaning of the number is that if the number for BOBL in chapter 1 is 418.57, a term is identified in every 418.57 words. Therefore, smaller number means higher indexing density. According to the figure, the density is higher in chapter 8 to 11 for BOBL, whereas the densities in all chapters for LINKL and TERML are similar to each other, except for the chapter 7. The pattern of BOBL is much different from those of LINKL and TERML, whereas the patterns of LINKL and TERML are similar with each other. In BOBL, the density is lower in chapter 1 to 4 and 13. Followings are the chapter titles in the book.

Ch 1. THE IMPORTANCE OF ACCESSING AND BROWSING INFORMATION AND COMMUNICATION; THE GENERAL APPROACH



〈Figure 4〉 Number of Words in Fulltext per Term

Ch 2. PERSPECTIVES ON ACCESS IN SIX RESEARCH LITERATURES
Ch 3. COMMON CONCEPTS ACROSS RESEARCH LITERATURES
Ch 4. UNIQUE ASPECTS ACROSS RESEARCH LITERATURES, AND A PRELIMINARY FRAMEWORK OF ACCESS
Ch 5. RESEARCH APPROACH
Ch 6. RESULTS: TESTING THE FRAMEWORK OF ACCESS
Ch 7. RESULTS: REFINING THE FRAMEWORK OF ACCESS
Ch 8. SUMMARY AND IMPLICATIONS OF THE FRAMEWORK OF ACCESS
Ch 9. PERSPECTIVES ON BROWSING IN SIX RESEARCH LITERATURES
Ch 10. A PRELIMINARY FRAMEWORK OF BROWSING
Ch 11. RESEARCH APPROACH
Ch 12. RESULTS: TESTING THE FRAMEWORK OF BROWSING
Ch 13. RESULTS: MOTIVATING THEMES AND PATTERNS OF BROWSING
Ch 14. RESULTS: A REFINED FRAMEWORK OF BROWSING
Ch 15. FUTURE RESEARCH AND IMPLICATIONS FOR THE FRAMEWORKS OF ACCESSING AND BROWSING INFORMATION AND COMMUNICATION

According the titles, chapter 1 to 4 are an introduction and literature reviews. In chapter 13, there are many examples from the authors' research data (interview data). The tendency of identifying

less number of terms in the introduction and literature reviews would be interpreted that the human indexer tends to identify unique terms, so that the terms from other literature and general terms regard as insignificant terms. The tendency of identifying less number of terms from examples and/or data for research would be interpreted that human indexers estimate the terms' significance as representatives of the document.

To compare among indexes, the terms in each index are matched to the terms in other 2 indexes as shown in <Table 4>. About 33% of terms identified by human (BOBL) are matched to terms in automatic index. That means only one-third terms by human are exactly same to terms in the text, and extracted; others are abstracted. LinkIt (38.36%) identifies more terms, which is identified by human, than Termer (28.88%). Similarity between terms by LinkIt and by Termer is expected to be high, but resulted moderate (54.03% for LIKNL → TERML and 64.69% for TERML → LINKL).

After examination of the statistics and the comparisons, the border between manual indexing and automatic indexing does still exist, however it becomes vague. And still which is better than the other is not answered. Lancaster (1998) said the purpose of indexing is to construct representations of published items. But the purpose can be varied in terms of its usage and its user. The index should be searchable and browsable (Anderson and Pérez-Carballo 2001a). Browsable displayed indexes with multi-term context-providing headings is common with manual indexing, but automatic indexing has many sophisticated techniques for selecting, combining, manipulating and weighting terms for searching. That is, manual index is more browsable than automatic index, and automatic is more searchable than manual. This comparison can be exactly employed in this study.

If the searching is performed by human, which index is better? An index by human is better than the automatic one, because the size of automatic index is much bigger than human can search. On the other hand, if an information retrieval system attempts to provide a browsing display, automatic index is better, because manual index is already coordinated so that the system cannot provide

〈Table 4〉 Results of matching terms in the indexes to each other

	Total number of matched terms	Average number of matched terms per chapter	Matching rate (matched terms /all terms)
BOBL → LINKL	352	23	38.36%
BOBL → TERML	270	18	28.88%
LINKL → BOBL	594	40	3.70%
LINKL → TERML	8,669	578	54.03%
TERML → BOBL	469	31	3.69%
TERML → LINKL	8,414	561	64.69%

enough terms to its users. Therefore, manual index is for the purpose to construct representations of published items for browsing items and for human use. On the contrary, automatic index is for the purpose to construct representations of published items for searching items and for machine use.

Then what is the usage of term identifiers, such as LinkIt and Termer, in terms of information retrieval? The one is that they can identify index terms for browsing and for human use. From the results of the comparison, they can partly identify significant terms from the text like a human indexer. The terms they identify are multi-word NP (LinkIt) with preposition (Termer), and occur frequently in the text (Termer). The terms are theoretically lexical and good candidates of multi-term context-providing headings, so that they can be used for browsing (Wacholder et al. 2001). Moreover, they have other advantages, such as keeping up with the volume of new text, new names, and terms, and efficiency.

The other is for searching and machine use. As described above, one of the characteristics of automatic indexing is exhaustive indexing with specific words (Anderson and Pérez-Carballo 2001a). Because the indexing unit is usually word in common indexing systems, their results have high recall but low precision. That is, they usually cannot identify lexical multi-word terms, so that they drop off the concepts in multi-word forms. Even though the words would be combined with and operator in a query, the user would also get unwanted information, because the locations of the words may not adjacent to each other in retrieved texts. In this case, the term identifiers are good alternatives as indexing systems, and they can support for high precision searching.

For the better use of the term identifier as an indexing system, some considerations should be discussed. First of all, more refined Natural Language Processing (NLP) techniques are essential. For an application of the algorithm with grammatical information, precise POS tagging technique is necessary. In fact, incomplete tagging performance of the pre-processing POS tagger, LinkIt and Termer cannot avoid producing inappropriate terms, and that is why a stopword list should be developed in the study. Another required traditional linguistic manipulation is truncation. Because POS information is important in the algorithms, removing prefixes, which can change the words' POS, such as *-ic, -ent,* and *-tion,* is not necessary. However, the truncation is still needed to make the forms of noun (singular-plural) consistently in an index. Refined NLP techniques should support to identify good index terms by the term identifiers.

Another consideration for a better performance of the term identifier is using document frequency in the algorithm. As described, the statistics said that human indexer tends not to identify terms from literature review part or an introduction part in the text. This can be interpreted that general terms or frequent terms in a domain are not good index terms not only for searching, but also for browsing. To identify and remove the general terms,

the old measure in information retrieval domain, inversed document frequency, would support. Also, human indexer can distinguish terms within contents of the text from terms within examples and/or research data. To distinguish the terms, more studies on document structures and on linguistic cue for the parts would be required.

5. Summary

To investigate the characteristics and differences of indexes by human and machine in terms of the term identification in a fulltext environment, a BOB index and two indexes from pseudo index systems (LinkIt and Termer) for a fulltext of a monograph are examined. LinkIt and Termer are not exact index systems, but multi-word term identifier for extracting significant terms from the text. In order to compare the terms in the indexes, some manipulation is employed in the indexes to produce terms only lists, because the formats of the three indexes are different from each other. After the manipulation, the term lists are abstracted to statistics, and are matched to each other.

Traditionally, manual index has strengths on browsability and vocabulary management, whereas automatic index has strengths on searchability, exhaustivity, and specificity. The contrast can exactly be utilized for the fulltext BOB index. Thus, manual indexing is for browsing and human use, whereas automatic indexing is for searching and machine use. But the border between them becomes vague, because the indexing system can identify multi-word terms as significant index terms, as human indexers do.

The term identifier would be useful for both purposes of indexes: browsing and searching. For browsing, the lexical, content-provided multi-word terms can be index terms directly. For searching, using the multi-word term can support to increase precision. However, there is a lot of room to be improved for the term identifiers. More linguistic manipulations, such as refined POS tagging and the truncation, should be employed. Also, using document frequency would be required for better identification of significant index terms in specific parts in a text, such as introduction and literature review parts. Moreover, this accompanies with additional research on document structures and on linguistic cues in the structures.

References

The American Society of Indexer. 2007. *The American Society of Indexers: Awards.* [cited 2008.5.23]. <http://www.asindexing.org/site/awards.shtml>.

The American Society of Indexer. 2006. *The American Society of Indexers: Indexing Evaluation Checklist: The index is the key to the book.* [cited 2008.5.23]. <http://www.asindexing.org/site/checklist.shtml>.

Anderson, J.D., & Pérez-Carballo, J. 2001a. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of manual indexing. *Information Processing and Management,* 37: 231-254.

Anderson, J.D., & Pérez-Carballo, J. 2001b. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing and Management,* 37: 255-277.

East, J. W. 2005. Subject retrieval of scholarly monographs via electronic databases. *Journal of Documentation,* 62: 597-605.

Evans, D.K. 1999. *A technical Description of the LinkIt System.* [cited 2003.1.14]. <http://www.columbia.edu/cu/cria/SigTops/LinkITTechDoc/>.

Gratch, B., Settel, B., & Atherton, P. 1978. Characteristics of book indexes for subject retrieval in the humanities and social sciences. *The Indexer,* 11: 14-23.

Hert, C.A., Jacob, E.K., & Dawson, P. 2000. A usability Assessment of online indexing structures in the networked environment. *Journal of the American Society for Information Science,* 51: 971-988.

Justeson, J.S., & Katz, S.M. 1994. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering,* 1: 9-27.

Lancaster, F.W. 1998. *Indexing and Abstracting in Theory and Practice.* Champaign, IL: University of Illinois, Graduate School of Library and Information Science.

Lathrop, L.M., Mauer, P. & Wyman, L.P. 1997. Quality and usability in indexes. In *Annual Conference Proceedings Society for Technical Communication.* [cited 2008.5.23]. <http://www.stc.org/ConfProceed/1997/PDFs/0124.pdf>.

Milstead, J.L. 1994. Needs for research in indexing. *Journal of the American Society for Information Science,* 45: 577-582.

Rasmussen, E.M. 1994. Indexing and retrieval from full-text. Introduction. In Fiden, R., Hahn, T.B., Rasmussen, E.M., & Smith, P.J. Eds.

Challenges in Indexing Electronic Text and Images. Medford, NJ: Learned Information. Inc., 241-245.

Rice, R.E., McCreadie, M., & Chang, S. 2001. *Accessing and Browsing Information and Communication: An Interdisciplinary Approach.* Cambridge, MA: MIT Press.

Sparck-Jones, K. 1973. Does indexing exhaustivity matter? *Journal of the American Society for Information Science,* 24: 313-316.

Wacholder, N. 1998. Simplex NPs clustered by head: A method for identifying significant topics within a document. In *Proceedings of the Workshop on the Computational Treatment of Nominals (COLING-ACL '98),* August 16, 1998. 70-79.

Wacholder, N., Evans, D.K., & Klavans, J.L. 2001. Automatic identification and organization of index terms for interactive browsing. *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries (JCDL '01),* June 24-28, 2001., Roanoke, Va: 126-134.

Wittman, C. 1990. Subheadings in Award-winning book indexes: a quantitative evaluation. *The Indexer,* 17: 3-6.

Yang, K. 2005. Information retrieval on the web. *Annual Review of Information Science and Technology,* 39: 33-80.

Appendix 1. Inappropriate word list

a	all	an	and	another	any
anything	as	did	do	each	everything
far	had	has	have	how	in
less	many	more	most	not	much
no	nothing	nowhere	of	one	one's
only	others	other	's	self	several
so	someone	something	such	that	the
these	thing	things	this	those	up
yet					