Evaluation of Automatically Identified Index Terms for Browsing Electronic Documents[1]

Nina Wacholder, Judith L. Klavans and David K. Evans
Columbia University
Department of Computer Science and
Center for Research on Information Access

1. Abstract

We present an evaluation of domain-independent natural language tools for use in the identification of significant concepts in documents. Using qualitative evaluation, we compare three shallow processing methods for extracting index terms, i.e., terms that can be used to model the content of documents. We focus on two criteria: quality and coverage. In terms of quality alone, our results show that technical term (TT) extraction [Justeson and Katz 1995] receives the highest rating. However, in terms of a combined quality and coverage metric, the Head Sorting (HS) method, described in [Wacholder 1998], outperforms both other methods, keyword (KW) and TT.

2. Introduction

In this paper, we consider the problem of how to evaluate the automatic identification of index terms that have been derived without recourse to lexicons or to other kinds of domain-specific information. By index terms, we mean natural language expressions that constitute a meaningful representation of a document for humans. The premise of this research is that if significant topics coherently represent information in a document, these topics can be used as index terms that approximate the content of individual documents in large collections of electronic documents.

We compare three shallow processing methods for identifying index terms:

- **Keywords (KW)** are terms identified by counting frequency of stemmed words in a document;
- **Technical terms (TT)** are noun phrases (NPs) or subparts of NPs repeated more than twice in a document [Justeson and Katz 1995];
- **Head sorted terms (HS)** are identified by a method in which simplex noun phrases (as defined below) are sorted by head and then ranked in decreasing order of frequency [Wacholder 1998].

The three methods that we evaluated are domain-independent in that they use statistical and/or linguistic properties that apply to any natural language document in any field. These methods are also corpus-independent, in that the ranking of terms for an individual document is not dependent on properties of the corpus.

2.1 Overview of methods and results

Subjects were drawn from two groups: professionals and students. Professionals included librarians and publishing professionals familiar with both manual and automatic text indexing. Students included undergraduate and graduate students with a variety of academic interests.

To assess terms, we used a standard qualitative ranking technique. We presented subjects with an article and a list of terms identified by one of the three methods. Subjects were asked to answer the following general question: "Would this term be useful in an electronic index for this article?" Terms were rated on a scale of 1 to 5, where 1 indicates a high quality term that should definitely be included in the index and 5 indicates a junk term that definitely should not be included. For example, the phrase *court-approved affirmative action plans* received an average rating of

[1] This research was partly funded by NSF IRI 97-12069, "Automatic identification of significant topics in domain independent full text documents" and NSF IRI 97-53054, "Computationally tractable methods for document analysis".

1 from the professionals, meaning that it was ranked as useful for the article; the KW *affirmative* received an average rating of 3.75, meaning that it was less useful; and the KW *action* received an average ranking of 4.5, meaning that it was not useful.

The goal of our research is to determine which method, or combination of methods, provides the best results. We measure results in terms of two criteria: **quality** and **coverage**. By quality, we mean that evaluators ranked terms high on the 1 to 5 scale from highest to lowest. By coverage, we mean the thoroughness with which the terms cover the significant topics in the document. Our methodology permits us to measure both criteria, as shown in Figure 4.

Our results from both the professionals and students show that TTs are superior with respect to quality; however, there are only a small number of TTs per document, so they do not provide adequate coverage in that they are not fully representative of the document as a whole. In contrast, KWs provide good coverage but relatively poor quality in that KWs are vague, and not well filtered. SNPs, which have been sorted using HS and filtered, provide a better balance of quality and coverage.

From our study, we draw the following conclusions:

- The KW approach identifies some useful index terms, but they are mixed in with a large number of low-ranked terms.
- The TT approach identifies high quality terms, but with low coverage, i.e., relatively few indexing terms.
- The HS approach achieves a balance between quality and coverage.

3. Domain-independent metrics for identifying significant topics

In order to identify significant topics in a document, a significance measure is needed, i.e., a method for determining which concepts in the document are relatively important for a given task. The need to determine the importance of a particular concept within a document is motivated by a range of applications, including information retrieval [Salton 1989], automatic determination of authorship [Mosteller and Wallace 1963], similarity metrics for cross-document clustering [Hatzivassiloglou et al. 1999], automatic indexing [Hodges et al. 1996] and input to summarization [Paice 1990].

For example, one of the earlier applications using frequency for identifying significant topics in a document was proposed by [Luhn 1958] for use in creating automatic abstracts. For each document, a list of stoplisted stems was created, and ranked by frequency; the most frequent keywords were used to identify significant sentences in the original document. Luhn's premise was that emphasis, as indicated by repetition of words and collocation is an indicator of significance. Namely, "the more often certain words are found in each other's company within a sentence, the more significance may be attributed to each of these words." This basic observation, although refined extensively by later summarization techniques (as reviewed in [Paice 1990]), relies on the capability of identifying significant concepts.

The standard IR technique known as tf*idf [Salton 1989] seeks to identify documents relevant to a particular query by relativizing keyword frequency in a document as compared to frequency in a corpus. This method can be used to locate at least some important concepts in full text. Although it has been effective for information retrieval, for other applications, such as human-oriented indexing, this technique is impractical. Ambiguity of stems (*trad* might refer to *trader* or *tradition*) and of isolated words (*state* might be a political entity or a mode of being) means that lists of keywords have not usually been used to represent the content of a document to human beings. Furthermore, humans have a difficult time processing stems and parts of words out of phrasal context.

The technical term (TT) method, another technique for identification of significant terms in text that can be used as index terms was introduced by [Justeson and Katz 1995], who developed an algorithm for identifying repeated multi-word phrases such as *central processing unit* in the computer domain or *word sense* in the lexical semantic domain. This algorithm identifies candidate TTs in a corpus by locating NPs consisting of

nouns, adjectives, and sometimes prepositional phrases. TTs are defined as those NPs, or their subparts, which occur above some frequency threshold in a corpus. However, as [Boguraev and Kennedy 1998] observe, the TT technique may not characterize the full content of documents. Indeed, even in a technical document, TTs do not provide adequate coverage of the NPs in a document that contribute to its content, especially since TTs are by definition multi-word. A truly domain-general method should apply to both technical and non-technical documents. The relevant difference between technical and non-technical documents is that in technical documents, many of the topics which are significant to the document as a whole may be also TTs.

[Wacholder 1998] proposed the method of Head Sorting for identifying significant topics that can be used to represent a source document. HS also uses a frequency measure to provide an approximation of topic significance. However, instead of counting frequency of stems or repetition of word sequences, this method counts frequency of a relatively easily identified grammatical element, heads of simplex noun phrases (SNPs). For common NPs (NPs whose head is a common noun), an SNP is a maximal NP that includes premodifiers such as determiners and possessives but not post-nominal constituents such as prepositions or relativizers. For example, *the well-known book* is an SNP but *the well-known book on asteroids* includes two SNPs, *well-known book* and *asteroids*. For proper names, an SNP is a name that refers to a single entity. For example, *Museum of the City of New York*, the name of an organization, is an SNP even though the organizational name incorporates a city name. Others, such as [Church 1988], have discussed a similar concept, sometimes called simple or base NPs.

The HS approach is based on the assumption that nominal elements can be used to convey the gist of a document. SNPs, which are semantically and syntactically coherent, appear to be at a good level of detail for content representation of the document. `

SNPs are identified by a system [Evans 1998; Evans et al. 2000] which sequentially parses text that has been tagged with part of speech using a finite state machine. Next, the complete list of SNPs identified in a document is sorted by the head of the phrase, which, at least for English-language common SNPs, is almost always the last word. The intuitive justification for sorting SNPs by head is based on the fundamental linguistic distinction between head and modifier: in general, a head makes a greater contribution to the syntax and semantics of a phrase than does a modifier. This linguistic insight can be extended to the document level. If, as a practical matter, it is necessary to rank the contribution to a whole document made by the sequence of words constituting an NP, the head should be ranked more highly than other words in the phrase. This distinction is important in linguistic theory; for example, [Jackendoff 1977] discusses the relationship of heads and modifiers in phrase structure. It is also important in NLP, where, for example, [Strzalkowski 1997] and [Evans and Zhai 1996] have used the distinction between heads and modifiers to add query terms to information retrieval systems.

Powerful corpus processing techniques have been developed to measure deviance from an average occurrence or co-occurrence in the corpus. In this paper we chose to evaluate methods that depend only on document-internal data, independent of corpus, domain or genre. We therefore did not use, for example, tf*idf, the purely statistical technique that is the used by most information retrieval systems, or [Smadja 1993], a hybrid statistical and symbolic technique for identifying collocations.

4. Experimental Method

To evaluate techniques, we performed a qualitative user evaluation in which the terms identified by each method were compared for usefulness as index terms.

4.1 Subjects

We performed our study with librarians, publishing professionals and undergraduate and graduate students at our university. 29 subjects participated in the study: 7 librarians and publishing professionals and 22 students.

4.2 Data

For this experiment, we selected three articles from the 1990 *Wall Street Journal* contained in the Tipster collection of documents. The articles were about 500 words in length.

To compare methods, each article was processed three times: 1) with SMART to identify stemmed keywords [Salton 1989]; 2) with an implementation of the TT algorithm based on [Justeson and Katz 1995]; and 3) with our implementation of the HS method. Output for one article is shown in Appendix A. Figure 1 shows the articles selected, their length in words and the number of index terms from each method for each article presented to the subjects.

DOC	words	KW	TT	HS
415-0109	509	63	4	49
516-0043	594	51	9	54
517-0062	514	52	8	57

Figure 1: Word and term count, by type, per article

The number of TTs is much lower than the number of KWs or HSs. This presented us with a problem: on the one hand, we were concerned about preserving the integrity of the three methods, each of which has their own logic, and at the same time, we were concerned to present lists that were balanced relative to each other. Toward this end, we made several decisions about presentation of the data:

1. **Threshold:** So that no bias would be unintentionally introduced, we presented subjects with all terms output by each method, up to a specified cut-off poin However, using lists of equal length for each method would have necessitated either omitting HSs and KWs or changing the definition of TTs. Therefore we made the following decisions:
   - For TTs, we included all identified terms;
   - For HSs, we included all terms whose head occurred more than once in the document;
   - For KWs, we included all terms in order of decreasing frequency, up to the point where we observed diminishing quality and where the number of KWs approximated the number of HSs.
2. **Order:** For the KW and TT approach, order is not significant. However, for the HS approach, the grouping together of phrases with common heads is, we claim, one of the advantages of the method. We therefore alphabetized the KWs and TTs in standard left to right order and alphabetized the HSs by head, e.g., *trust account* precedes *money market fund*.
3. **Morphological expansion:** The KW approach identifies stems which represent a set of one or more morphological variants of the stem. Since in some cases the stem is not an English word, we expanded each stem to include the morphological variants that actually occurred in the article. For example, for the stem *reject*, we listed *rejected* and *rejecting* but did not list *rejects*, which did not occur in the article.

4.3 Presentation to subjects

Each subject was presented with three articles. For one article, the subject received a head sorted list of HSs; for another article, the subject received a list of technical terms, and for the third article, the subject saw a list of keywords. No time limit was placed on the task.

5. Results

Our results for the three types of terms, by document, are shown in Figure 2. Although we asked subjects to rate three articles, some volunteers rated only two. All results were included.

Doc	Avg KW rating	Avg TT rating	Avg HS rating
900405-0109	3.08	1.45	2.71
900516-0043	3.73	2.19	2.71
900517-0062	2.98	1.7	3.25
Avg of Avgs	**3.27**	**1.79**	**2.89**

Figure 2: Average ratings of 3 types of index terms

5.1 Quality

For the three lists of index terms, TTs received the highest ratings for all three documents—an average of 1.79 on the scale of 1 to 5, with 1 being the best rating. HS came in second, with an average of 2.89, and KW came in last with an average of 3.27. It should be noted that averaging the average conceals the fact that the number of TTs is much lower than the other two types of terms, as shown in Figure 1.

Figure 3 (included before Appendix A) shows cumulative rankings of terms by method. The X axis represents ratings awarded by subjects. The Y axis reflects the percentage of terms receiving a given rank or better. All data series must reach 100% since every term has been assigned a rating by the evaluators. At any given data point, a larger value indicates that a larger percentage of that series' data has that particular rating or better. For example, 100% of the TTs have a rating of 3 or better; while only about 30% of the terms of the lowest-scoring KW document received a score of 3 or better. In two out of the three documents, HS terms fall between TTs and KWs.

5.2 Coverage

The graph in Figure 3 shows results for quality, not coverage. In contrast, Figure 4, which shows the total number of terms rated at or below specified rankings, allows us to measure quality and coverage. (1 is the highest rating; 5 is the lowest.) This figure shows that the HS method identifies more high quality terms than the TT method does.

	Number of terms ranked at or better than			
Method	**2**	**3**	**4**	**5**
KW	27	75	124	166
HS	41	96	132	160
TT	15	21	21	21

Figure 4: Running total of terms identified at or below a specified rank

TT clearly identifies the highest quality terms: 100% of TTs receive a rating of 2 or better. However, only 8 TTs received a rating of 2 or better (38% of the total), while 41 HSs received a rating of 2 or better (26% of the total). This indicates that the TT method misses many high quality terms. KW, the least discriminating method in terms of quality, also provides better coverage than does TT.

This result is consistent with our observation that TT identifies the highest quality terms, but there are very few of them: an average of 7 per 500 words compared to over 50 for HS and KW. Therefore there is a need for additional high quality terms. The list of HSs received a higher average rating than did the list of KWs, as shown in Figure 2. This is consistent with our expectation that phrases containing more content-bearing modifiers would be perceived as more useful index terms than would single word phrases consisting only of heads.

5.3 Ranking variability

The difference in the average ratings for the list of KWs and the list of head-sorted SNPs was less than expected. The small difference in average ratings for the HS list and the KW list can be explained, at least in part, by two factors: 1) Differences among professionals and students in inter-subject agreement and reliability; 2) A discrepancy in the rating of single word terms across term types.

22 students and 7 professionals participated in the study. Figure 5 shows differences in the ratings of professionals and of students.

	Professionals	**Students**
KW	2.64	3.30
HS	2.3	3.03
TT	1.49	2.1

Figure 5: Average ratings, by term type, of professionals and students

When variation in the scores for terms was calculated using standard deviation, the standard deviation for the professionals was 0.78, while for the students it was 1.02. Because of the relatively low number of professionals, the standard deviation was calculated only over terms that were rated by more than one professional. A review of the students' results

showed that they appeared not to be as careful as the professionals. For example, the phrase 'Wall Street Journal' was included on the HS list only because it is specified as the document source. However, four of the eight students assigned this term a high rating (1 or 2); this is puzzling because the document is about asbestos-related disease. The other four students assigned a 4 or 5 to 'Wall Street Journal', as we expected. But the average score for this term was 3, due to the anomalous ratings. We therefore have more confidence in the reliability of the professional ratings, even though there are relatively few of them.

We examined some of the differences in rating for term types. Single word index terms are rated more highly by professionals when they appear in the context of other single word index terms, but are downrated in the context of phrasal expansions that make the meaning of the one-word term more specific. The KW list and HS list overlap when the SNP consists only of a single word (the head) or only of a head modified by determiners. When the same word appears in both lists in identical form, the token in the KW list tends to receive a better rating than the token does when it appears in the HS list, where it is often followed by expansions of the head. For example, the word *exposure* received an average rating of 2.2 when it appeared on the KW list, but a rating of only 2.75 on the HS list. However, the more specific phrase *racial quotas*, which immediately followed *quota* on the HS list received a rating of 1.

To better understand these differences, we selected 40 multi-word phrases and examined the average score that the phrase received in the TT and HS lists, and compared it to the average ratings that individual words received in the KW list. We found that in about half of the cases (21 of 40), the phrase as a whole and the individual words in the phrase received similar scores, as in Example 1 in Figure 6. In just over one-fourth of the cases (12 of 40), the phrase scored well, but scores from the individual words were rated from good to poor, as in Example 2. In about one-eighth of the cases (6 of 40), the phrase scored well, but the individual words scored poorly, as in Example 3. Finally, in only one case, shown in Example 4 of Figure 6, the phrase scored poorly but the individual words scored well.

	Phrase	Word 1	Word 2
1	Supreme Court (1.5)	Supreme (1)	Court (1.25)
2	reverse discrimination (1)	reverse (3.25)	discrimination (3.25)
3	lymph system (1)	lymph (1)	system (5)
4	employment decisions (2.75)	employment (1.25)	decisions (1.25)

Figure 6: Comparison of scores of phrases and single words

This shows that single words in isolation are judged differently than the same word when presented in the context of a larger phrase. These results have important implications in the design of indexing tools.

6. Conclusion

Our results show that the head sorting technique outperforms two other indexing methods, technical terms and keywords, as measured by balance of quality and coverage. We have performed a qualitative evaluation of three techniques for identifying significant terms in a document, driven by an indexing task. Such an application can be used to create a profile or thumbnail of a document by presenting to users a set of terms which can be considered to be a representation of the content of the document. We have used human judges to evaluate the effectiveness of each method. This research is a contribution to the overall evaluation of computational linguistic tools in terms of their usefulness for human-oriented computational applications.

8. References

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

Church, Kenneth Ward (1988) "A stochastic parts program and noun phrase parser for unrestricted text", in *Proceedings of the Second*

Conference on Applied Natural Language Processing, pp.136-143.

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. University of California, Santa Cruz, California, Morgan Kaufmann Publishers.

Evans, David K. (1998) LinkIT Documentation, Columbia University Department of Computer Science Report.

Evans, David K., Klavans, Judith, and Wacholder, Nina (2000) "Document processing with LinkIT", RIAO Conference, Paris, France, to appear.

Hatzivassiloglou, Vasileios, Judith L. Klavans and Eleazar Eskin (1999) "Detecting text similarity over short passages: exploring linguistic feature combinations via machine learning", *Proceedings of the EMNLP/VLC-99 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, University of Maryland, College Park, MD.

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.

Jackendoff, Ray (1977) *X-bar Syntax: A Study of Phrase Structure*, MIT Press, Cambridge, MA.

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.

Luhn, Hans P. (1958) "The automatic creation of literature abstracts", *IBM Journal*, 159-165.

Mosteller, Frederick and David L. Wallace (1963) "Inference in an authorship problem", *Journal of the American Statistical Association* 58(302):275-309. Available at http://www.jstor.org/.

Paice, Chris D. (1990) "Constructing literature abstracts by computer: techniques and prospects". *Information Processing & Management* 26(1):171-186.

Salton, Gerald (1989) *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.

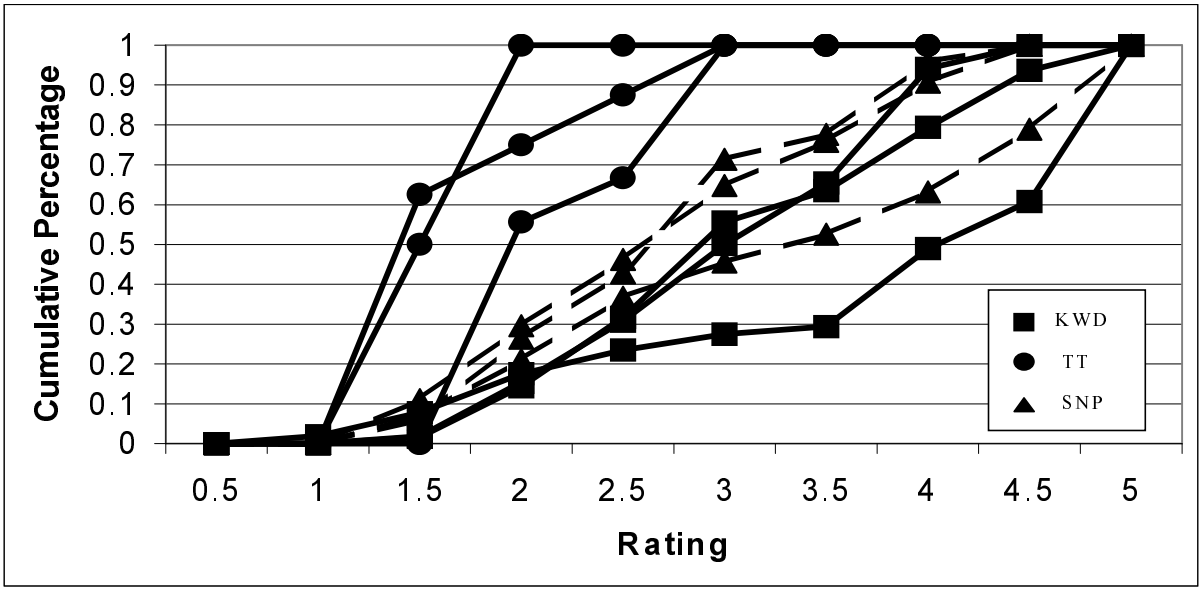
Smadja, Frank (1993) "Retrieving collocations from text", *Computational Linguistics* 19(1):143-177.

Strzalkowski, Tomek (1997) "Building effective queries in natural language information retrieval", *Proceedings of the ANLP*, ACL, Washington, DC., pp.299-306.

Wacholder, Nina (1998) "Simplex NPS sorted by head: a method for identifying significant topics within a document", *Proceedings of the Workshop on the Computational Treatment of Nominals*, pp.70-79. Montreal, Canada.

Figure 3: Cumulative ranking of terms, by method

Appendix A: Terms identified in WSJ900405-0109



HSs

amendments
Hatch amendment
other amendments
attempts
bias
job bias
intentional bias
bill
committee
Senate labor Committee
court
Supreme Court
co-workers
decisions
Supreme Court decisions
employment decisions
Democrats
discrimination
reverse discrimination
employees
women employees
employers
groups
civil-rights groups
conservative policy groups
Orrin Hatch
health
discriminatory impact
Job-Bias Measure
basic employment anti-discrimination law
1866 civil-rights law
lawsuits
lawmakers
legislation
comprehensive legislation
more modest measure
minority/minorities
panel
plans
court-approved affirmative action plans
discriminatory seniority plans
practices
employment practices
quotas
racial quotas
right/rights
equal rights
year

Keywords

action
address/addressing
adopt/adopted
affirmative
agree
aimed
alleged/alleging
amend
approved
attempt/attempts
bias
bill
Bush
challenge
circumstances
civil
clears
committee
court/Court
decision
Democrats
discrimination
employment/employers/employees
force/Force
give/giving
GOP
groups
Hatch
health
high
impact
job
justify
labor/Labor
law
lawmakers
lawsuits
legislative/legislation
make
measure
minority/minorities
Mr.
overturning
panel
plans
policy
practices
quotas
racial
rejected/rejecting
reverse
rights
rules/ruling
safety
Sen./Sens.
Senate
shown
street
Supreme; vote/voted
women
workers
year

Technical terms

discriminatory impact
employment practice
Senator Hatch
Supreme Court