Thesaurus Based Term Ranking for Keyword Extraction

Luit Gazendam
Novay
Enschede, The Netherlands
Luit@gazendam.net

Christian Wartena
Novay
Enschede, The Netherlands
Christian.Wartena@novay.nl

Rogier Brussee
University of Applied Sciences Utrecht
Utrecht, The Netherlands
Rogier.Brussee@hu.nl

Abstract—**In many cases keywords from a restricted set of possible keywords have to be assigned to texts. A common way to find the best keywords is to rank terms occurring in the text according to their tf.idf value. This requires a corpus of texts from which document frequencies can be derived. In this paper we show that we can obtain results of the same quality without the usage of a background corpus, using relations between terms provided in a thesaurus.**

I. Introduction

As a growing number of documents is stored and is electronically available, a good annotation of these documents is necessary to enable efficient retrieval. Keyword annotation of documents is an important feature for document retrieval, classification, topic search and other tasks even if full text search is available. Keywords provide a concise and precise high-level summarization of a document.

We distinguish two variants of annotation with keywords. In the first variant annotators can freely choose the keywords that describe the document. In the second flavor the keywords have to be taken from a restricted vocabulary or thesaurus. The second approach has a lot of advantages since it guarantees a certain degree of consistency. Disadvantages are clearly that a thesaurus has to be constructed and maintained and that it might be the case that no term matching the content of the document is available.

Manual annotation of documents with keywords is a tedious task. Automatic keyword extraction from documents therefore seems to be an important tool. Automatically extracted keywords may either be directly assigned to documents or suggested to human annotators. Basically we can divide various approaches to automatic keyword extraction into two main ways of thinking. In the first approach there is a relatively small numbers (usually in a range from a dozen up to a few hundreds) of keywords and keyword assignment is treated as classification. The second approach tries to identify words in the text that are important and characteristic for that text. While the former approach uses a restricted vocabulary by definition, the latter is usually associated with freely chosen keywords. In the present paper we study the extraction of keywords from texts but still using a restricted vocabulary.

There are two main reasons to combine extraction from keywords from texts with a controlled vocabulary, apart from the fact that many archives and libraries have thesauri that they want to continue using. In the first place, thesauri used for annotation may contain up to 30,000 terms and more. This size becomes problematic for classification methods, since usually not enough training data is available for each category. Moreover, results for classification decline with a growing number of classes. The second reason to use a thesaurus in combination with keyword extraction is that a thesaurus is also a large knowledge base on the domain under consideration. This knowledge can be exploited for keyword extraction. Usually a corpus of texts is needed to train a keyword extraction or classification algorithm or to determine the importance of a word in the whole collection relative to the importance in a document. Below we will show some results in which this analysis of a background corpus is replaced by the analysis of a thesaurus from which the importance of a term also can be determined. This makes the results of keyword extraction independent from the collection chosen for training and comparison.

The organization of this paper is as follows. In section II we discuss related work. In III we present three ways of thesaurus based keyword extraction. In IV we describe an experiment to compare these methods and in V we finally discuss the results.

II. Related Work

As ranking is one of the central issues in information retrieval there is a vast literature on term weighting. In the article by Salton[1], extensive experiments with different weight structures are reported and it is mentioned that the right weight structure is a central area of research since the late 1960's. See also the 1972 article reprinted as [2] where weighting for specificity of a term based on $1 + log(\#\text{documents}/\#\text{term occurrences})$ is already proposed based on empirical data. This term weighting is subsequently refined together with Robertson [3], studied in the light of latent semantic analysis in [4], given a detailed statistical analysis in [5], and a probabilistic interpretation in [6].

More closely related to the work presented in this paper are various studies to improve automatic keyword extraction/suggestion with information stored in thesauri. Although Kamps[7] questions the value of thesauri for annotation, he proposes a term weighting and ranking strategy based on thesaurus relations for cross lingual retrieval of documents with good results on the CLEF 2002 data set. Hulth et al.[8] and Medelyan and Witten[9], [10] do show an improvement in automatic keyword extraction with machine learning techniques using thesaurus relations. However, we could not reproduce these effects on the dataset used for the work described below. A second approach is by [11] who do not train their model on existing data and only use information from the thesaurus in combination with Bayesian statistics and probabilistic inferencing to suggest keywords. In the same manner [12] only use thesaurus information and use PageRank to determine the most central WordNet keywords in the graphs which could be constructed with the WordNet relations between the keywords appearing in the text. We also were not able to reproduce these results for our dataset.

III. Annotation and Ranking

Our approach to automatically suggesting keywords is based on information extraction techniques applied to textual resources. Our system transforms these texts into a suggestion list of thesaurus keywords. The system consists of two parts: a *text annotator* which identifies all terms occurring in the text, and a *ranking process* which transforms the set annotations into ranked lists.

The text annotator scans a text for all possible textual representations of concepts related to thesaurus terms, and annotates all different lexical occurrences of a concept with its Unique Resource Identifier (URI)[1]. For this task we used Apolda [13][2], a plug-in for GATE [14] and UIMA [15]. As input Apolda uses a lexicalized ontology, which contains for each concept multiple lexical representations, such as preferred spelling, plural spelling, synonyms and annotates the terms occurring in the text with the corresponding URI. Apolda searches for representations matching the words in the text or their lemmata.

A. Ranking Weights of thesaurus terms

In order to rank the thesaurus terms for a document, we assign weights to each term The *tf.idf* measure uses a corpus of documents as a frame of reference. The *tf.rr* measure uses only the thesaurus as frame of reference.

The *tf.idf* of a term in a document depends on the frequency of the term in the document and on the number of documents in the collection containing that term. We use the following standard variant of *tf.idf* [16, p544]:

$$tf.idf(t,d) = n(d,t) \log \frac{N}{df(t)}$$

where df is the number of documents d' for which $n(d',t) > 0$, $n(d,t)$ is the number of occurrences of t in d and N is the number of documents in the corpus. Note that this is not simply the *tf.idf* value of words in the text, but applied URIs discovered in the annotation phase.

A rich source of information for determining the importance of a term in a text is provided by all other terms present in that text. The basic idea is that the importance of a term is not only apparent from the number of occurrences of that term but also from the number of related terms in the text. We can use the relations specified in a thesaurus to find relations between terms in the text. To avoid confusion, in the following we will refer to the relations in the text as realized relations, i.e. relations that are specified in the thesaurus and for which terms realizing the terms in the relation are both found in the text. From the number of realized relations we can compute a term weight. Obviously, there are several ways to do this.

To be precise, we construct the set of all terms represented the text. We then construct a graph with the terms from this set as nodes and two types of edges. The "distance 1" edges are all the relations between the nodes as in the thesaurus. Here we make no distinction between the type of relation (like broader term, related term, etc.). For the "distance 2" edges we take all relations that can be composed from two thesaurus relations. An example of such a graph is given in Fig. 1. The intermediate terms of the composed relations that are formally not part of the graph are shown in this picture as well. Note that for the number of realized relations we do not take the number of instances of the terms into account. To compute the overall weight of a term t in document d we multiply term frequency with a weight that is determined by the number of realized relations at distance 1 ($r_1(t,d)$) and distance 2 ($r_2(t,d)$). This gives us the following formula.

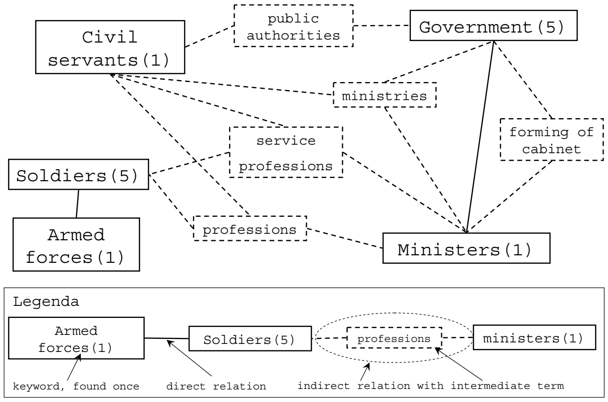

Fig. 1. Relations found between a set of keywords

$$tf.rr(t,d) = tf(t,d)rr(t,d) \tag{1}$$

where

$$\begin{aligned} tf(t,d) &= 1 + \log(n(t,d)) \\ rr(t,d) &= 1 + \mu r_1(t,d) + \mu^2 r_2(t,d) \end{aligned}$$

with $n(t,d)$ the number of occurrences of t in d, $\mu = \alpha/\text{avlinks}$ and where avlinks is the average number of relations a term has in the thesaurus (the out degree). This average number of links determines the average number of reachable thesaurus terms. At distance 1 this number of reachable terms

[1] this URI identifies both the concept and the Knowledge Organization System from which it stems

[2] http://apolda.sourceforge.net/

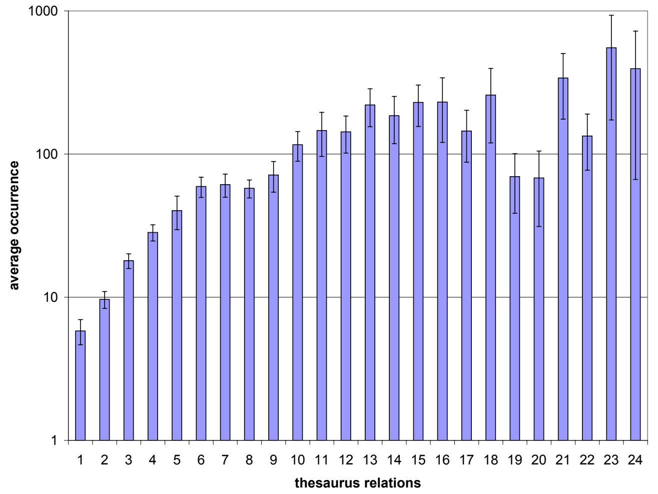

Fig. 2. Number of occurrences in the corpus for keywords with N thesaurus relations

is avlinks. At distance 2 this number is avlinks^2. The factor α is a damping factor which reduces the contribution of distance 2 relations compared to distance 1 relations. We set $\alpha = 1/2$ from which we expect that the contribution of distance 2 relations to *tf.rr* is about $1/2$ that of the distance 1 relations. This is also found in practice.

The proposed weighting scheme captures the idea that central (and thus important) terms are related to other concepts mentioned in the text. The results of the weighting scheme, however, also depend on the degree to which a concept is worked out. We here find two tendencies: in the first place, terms unimportant for the domain, usually are not worked out very well and therefore do not have many relations in the thesaurus. These terms therefore will never score very high. This effect might contribute positively to filter out unlikely candidates. However, we usually also find a number of terms with an extremely huge number of relations. These terms often serve to tie together some distinct subparts of the thesaurus. Thus their number of relations does not correspond to their importance as keyword. Figure 2 shows the relation between the number of specified relations in the thesaurus and the number of times they are used as keyword in the corpus presented in the next section.

One advantage of this weighting over *tf.idf* is that a term occurring in the text that is very specific but not related to the main subject of the text will not get a high rating. The other effect of *tf.idf*, the suppression of common words is not important in our scenario, since we restrict to thesaurus terms anyway. The other main advantage over *tf.idf* is that *tf.rr* is independent of a corpus and depends only on the document and the thesaurus.

IV. Experiment

The experiments were conducted at the Netherlands Institute for Sound and Vision, which is in charge of archiving publicly broadcasted TV and radio programs in the Netherlands. Their cataloguers annotate the audiovisual broadcast according to strict guidelines. During annotation cataloguers consult the audio visual material and often also consult available contextual information such as TV-guide synopses, official TV-programs web site texts and subtitles. All catalogue descriptions conform to a metadata scheme called iMMiX, which is an adaptation for 'audiovisual' catalogue data of the FRBR data model[3] developed by the international federation of library associations (IFLA). Choices for some of the iMMiX fields (subject, location, persons etc.) are restricted to a thesaurus called GTAA.

A. Material

For our experiments we created a corpus of 258 broadcasted TV-documentaries, 80% of which belonged to three series of TV-programs: *Andere Tijden*, a series of Dutch historical documentaries, *Beeldenstorm*, a series of art documentaries and *Dokwerk*, a series of historical political documentaries.

Each broadcast has context documents in the form of one or more texts from the broadcasters web site. The 258 TV-broadcasts are associated with 362 context documents which varied in length between 25 and 7000 words with an average of 1000 words.

Each program also has a catalogue description created manually by cataloguers from Sound and Vision. Each description contains at least 1 and at most 15 keywords with an average of 5.7 and a standard deviation of 3.2 keywords. These keywords are the ground truth against which we evaluate the *tf.idf* baseline and the two other ranking algorithms in the experiments.

The GTAA (a Dutch acronym for "Common Thesaurus [for] Audiovisual Archives") is constructed over the last 15 years and is updated bi-weekly by information specialists. It adheres to the ISO 2788 guidelines for the establishment and development of monolingual thesauri[17]. It contains about 160000 terms, organized in 6 facets: **Locations, People, Names (of organizations, events etc.), Makers, Genres** and **Subjects**. This latest facet contains 3860 keywords and 20 591 relations between the keywords belonging to the relationships of Broader Term, Narrower Term, Related Term and Use/Use for. It also contains linguistic information such as preferred textual representations of keywords and non-preferred representations. Each keyword on average has 1 broader, 1 narrower and 3.5 related terms. Cataloguers are instructed to select keyword that describe the program as a whole, that are specific and that allow for good retrieval.

Apolda requires that a thesaurus is represented in the SKOS data model [18]. We therefore used a version of the thesaurus which is transformed to SKOS using a standard approach[19]. Subsequently we enriched this SKOS version with singular forms.

[3] Functional Requirements for Bibliographical Record, www.ifla.org/VII/s13/frbr/frbr.pdf (06/03/09)

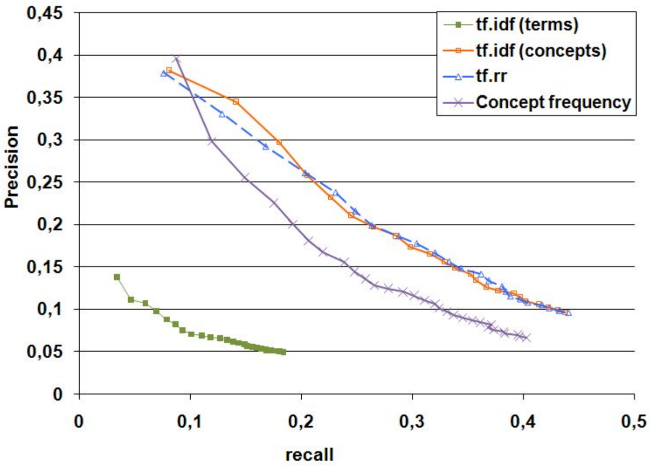

Fig. 3. Precision-recall graph for three ranking algorithms

B. Evaluation against manually assigned keywords

In our experiments we generate and rank keyword suggestions for TV-programs from contextual resources, and we evaluate these against manually assigned keywords. Two factors are problematic during this evaluation: 1) our automatic process is not performing exactly the same task as cataloguers since the automatic process ignores the audio-visual material, and 2) cataloguers hardly agree among one another when assigning keywords.

In practice using only context documents can in fact be an advantage: when contextual information is available it often summarizes the content of the program which makes it easier to find summarizing keywords.

When a controlled vocabulary is used, typical measures of inter-cataloguer consistency are in the 13% to 77% range with an average of 44% [20]. Analysis of this disagreement shows that a significant portion of these differences are in fact small semantic differences. Such differences are mostly irrelevant for the intended usage of the keywords, but can be problematic when manual annotations serve as a gold standard for the evaluation of our automatic annotation suggestions.

V. Results

The graph in Figure 3 displays the precision recall curves for the three different rankings: *tf.idf* based on concepts as discussed above, *tf.idf* based on words and *tf.rr*. For computing the word based *tf.idf*-value we lemmatized all words as in the other variants but we did not restrict the statistics to words mentioned in the thesaurus. For evaluation, different forms with respect to number (singular/plural) and other morphological or spelling variants were counted as true positives.

From this figure we directly see the importance of using the thesaurus in the keyword selection process. In the first place irrelevant terms are filtered out. Thus the precision is increased significantly. In the second place the recall is improved, since a lot of non-preferred terms are taken into account. These terms are detected by Apolda and mapped to their canonical form according to the GTAA ontology.

TABLE I
The suggested terms for Andere Tijden 2003-11-11: Mining disaster at Marcinelle

rank	*tf.idf*	*tf.rr*	Catalogue
1	**miners**	**disasters**	**history**
2	**disasters**	**miners**	**foreign employees**
3	fire	fire	**disasters**
4	cables	*fires*	**coal mines**
5	**foreign employees**	**foreign employees**	**miners**
6	*lignite*	*lignite*	
7	safety	*immigrants*	
8	governments	*fire brigade*	
9	*fire brigade*	families	
10	*fires*	governments	
11	elevators	*mining*	
12	*immigrants*	safety	
13	law	**coal mines**	
14	engineers	**history**	

The second, and most important conclusion is, that the usage of thesaurus relations in the text can compensate completely the absence of a reference corpus.

A qualitative analysis of the lists generated by the three different ranking algorithms for one specific case may give us some more insight into the algorithms' qualities and shortcomings. The TV-program *Andere Tijden 11-11-2003, Mining accident at Marcinelle* is chosen for this illustration.

Sound and Visions' catalogue describes this program as follows: Episode of the weekly program "Andere Tijden", in which a mining accident in the fifties of last century in Belgium is addressed. In this mining accident many Italian foreign workers died in a fire. The first 14 ranks generated by our four settings are displayed in table V. The cataloguer attached the keywords *history, disasters, coal mines, miners* and *foreign employees* to this program. Note that another cataloguer on average would have chosen only 2.5 of the same keyword and 2.5 other keywords to describe this program. The catalogue keywords are not ranked (all are deemed equally correct).

The keywords in **boldface** are exact matches with the catalogue keywords. The keywords in *italics* are semantically correct at distance 1 and the keywords in normal font are wrong.

While this is only a single example, the table suggests some observations. First we see that in terms of exact matches each list contains the three correct suggestions **miners**, **disasters** and **foreign employees** among the first 5. The *tf.rr* has the two other catalogue keywords **coal mines** and **history** at rank 13 and 14.

The second observation is that the *tf.rr* has the most distance 2 matches (semantically correct, but not exact suggestions) in the list: *fires, lignite, immigrants, fire brigade* and *mining*, while *tf.idf* has as a consequence more terms that are evaluated as incorrect. Some of these as incorrect evaluated terms, such as *fire (Dutch: vuur)* seem quite reasonable, but in the GTAA this means the concept or element of fire. A *fire* is referred to with the plural *fires (Dutch: brand)*, which is semantically correct as it has a relation to **disasters**.

The final observation is that the two methods seem to have different measures of coherence between the suggestions. The *tf.rr* seems the most coherent (it has *fires* at the fourth rank compared to *cables* in *tf.idf*). The use of relations among the found keywords creates this coherence. This element of coherence may be pleasant for catalogers receiving annotation suggestions.

The observations made form studying one example in detail suggest, that, while both methods perform equally well using a formal evaluation method, the results of the *tf.rr* method are slightly better in some aspects. The same picture also arises from other examples.

VI. Conclusions

In this paper we study the extraction and ranking of keywords with a restricted vocabulary. Although we use the vocabulary both for the annotation of parts of text with thesaurus term URI's and for the ranking of these thesaurus annotations, the focus lies on the value of the thesaurus for the ranking. The idea behind the thesaurus based ranking is that a thesaurus is a large knowledge base on the domain under consideration which can be used automatically.

We developed a new weighting scheme for ranking words to be used as keywords. This new scheme, *tf.rr*, uses both the frequency information of a term in a document and the number of realized thesaurus relations between the thesaurus terms found in the specific document, but thus not need any kind of training or statistics from a reference corpus.

In an experiment we compared ranked lists of suggestions against manually assigned keywords at the Netherlands Institute for Sound and Vision. We implemented a semantic evaluation next to the classic evaluation to tackle the problem of inter annotator disagreement during evaluation against the manually assigned keywords. Our results showed that the new weighting scheme performs equally well as the classical *tf.idf*. This suggests that the usage of thesaurus relations can replace the usage of a reference corpus.

A qualitative inspection of the results suggests that the coherence between the suggestions seems bigger for the *tf.rr* algorithm. This is important when we use the algorithm to suggest keywords to catalogers, a very probable practical implementation of an automatic keyword extraction algorithm within working archives.

VII. Acknowledgments

This research was funded by the European Commission FP7 project MyMedia (http://www.mymediaproject.org) under the grant agreement no. 215006 and by the NWO CHOICE@CATCH project no. 640.001.402. The authors would like to thank Rogier Brussee for many helpfull comments.

References

[1] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval." Cornell University, Tech. Rep., 1987. [Online]. Available: http://hdl.handle.net/1813/6721

[2] K. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 60, pp. 493–502, 2004.

[3] S. Robertson and K. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, vol. 27, no. 3, 1976.

[4] S. Dumais, "Improving the retrieval of information from external sources," *Behavior Research Methods, Instruments and Computers*, vol. 23, no. 2, pp. 229–236, 1991.

[5] W. Greiff, "A theory of term weighting based on exploratory data analysis," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, 1998, pp. 11–19.

[6] D. Hiemstra, "A probabilistic justification for using tf× idf term weighting in information retrieval," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 131–139, 2000.

[7] J. Kamps, "Improving retrieval effectiveness by reranking documents based on controlled vocabulary," in *Advances in Information Retrieval: 26th European Conference on IR Research (ECIR 2004)*, ser. Lecture Notes in Computer Science, S. McDonald and J. Tait, Eds., vol. 2997. Springer-Verlag, Heidelberg, 2004, pp. 283–295.

[8] A. Hulth, J. Karlgren, A. Jonsson, and L. Bostrom, H. an Asker, "Automatic keyword extraction using domain knowledge," *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, 2004.

[9] O. Medelyan and I. H. Witten, "Thesaurus-based index term extraction for agricultural documents," in *Proc. of the 6th Agricultural Ontology Service workshop*, 2005.

[10] O. Medelyan and I. Witten, "Thesaurus based automatic keyphrase indexing," in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. ACM New York, NY, USA, 2006, pp. 296–297.

[11] L. M. De Campos, J. M. Fernández-Luna, J. F. Huete, and A. E. Romero, "Automatic indexing from a thesaurus using bayesian networks," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, K. Mellouli, Ed. LNCS 4724, Springer, 2007, pp. 865–877.

[12] J. Wang, J. Liu, and C. Wang, "Keyword extraction based on pagerank," *Advances in Knowledge Discovery and Data Mining*, vol. 4426, pp. 857–864, 2007.

[13] C. Wartena, R. Brussee, L. Gazendam, and W. Huijsen, "Apolda: A practical tool for semantic annotation," in *The 4th International Workshop on Text-based Information Retrieval (TIR 2007)*, Regensburg, Germany, September 2007.

[14] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A framework and graphical development environment for robust NLP tools and applications," in *Proceedings of the 40th Anniversary Meeting of the ACL*, 2002.

[15] D. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment," *Natural Language Engineering*, vol. 10, no. 3-4, pp. 327–348, 2004.

[16] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, 1999.

[17] ISO, "Guidelines for the establishment and development of monolingual thesauri," *ISO 2788-1986*, 1986.

[18] A. Miles and D. Brickley, "SKOS core guide," World Wide Web Consortium, W3C Working Draft, November 2005, electronic document. Accessed February 2008. Available from: http://www.w3.org/TR/swbp-skos-core-guide/.

[19] M. van Assem, V. Malaise, A. Miles, and G. Schreiber, "A method to convert thesauri to skos," in *Proceedings of the Third European Semantic Web Conference (ESWC'06)*, ser. Lecture Notes in Computer Science, no. 4011, Budva, Montenegro, June 2006, pp. 95–109. [Online]. Available: http://www.cs.vu.nl/~mark/papers/Assem06b.pdf

[20] K. Leininger, "Inter-indexer consistency in psycinfo," *Journal of Librarianship and Information Science*, vol. 32, no. 1, pp. 4–8, 2000.