

COMPARATION OF TITLE INDEXING AND ABSTRACT INDEXING IN INFORMATION RETRIEVAL SYSTEM

Indra Budi and Zainal A. Hasibuan
Faculty of Computer Science, University of Indonesia
Kampus UI Depok 16424 Indonesia
Phone/Fax: (021)7863419/(021)7863415
Email: indra@cs.ui.ac.id, zhasibua@cs.ui.ac.id

Abstract

We could develop index based on each part of document like title, author, publisher, abstract, document content and combination among those parts. Naturally, the size of index based on title of document (title indexing) smaller than index based on abstract of document (abstract indexing). So, retrieval process on title indexing faster than abstract indexing, but we could not know the effectiveness of the result. Objective of this research is to determine effectiveness indexing based on title of document (title indexing) compared indexing based on abstract of document (abstract indexing). We use 1162 abstracts of document about research on science and nuclear technology from: BATAN as our corpus. Our experiment showed that the average percentage of the number activated document based on title indexing is less than abstract indexing (7.76% : 17.61%). The average of relevant document based on title indexing just a bit less than abstract indexing (14% : 15%).

Keyword: indexing, title indexing, abstract indexing

1. INTRODUCTION

In Information Retrieval System (IRS) area, document usually can be divided into three parts, which are: title, abstract and content. Naturally, title is the smallest part of the document that consist of one or maximum two sentences. Each part of the document can be used as basis of document index. Indexing (document representation) is an important part of the information storage and retrieval process [4]. Index provide a vital link between stored document data and its later retrieval.

The size index may have implication to the capacity of storage and retrieval time. If the number of terms as index increasing then we will need more capacity of storage and more time for retrieval. We have been known that the number terms in title of document smaller than abstract and the document content. So, if we used title as indexing then we would save more storage and would save more retrieval time. Our corpus [12], showed that title indexing decreased the number of index term 4.4 times than abstract indexing (2062:9254).

But the question is arisen, what would the effectiveness of the retrieval? If we know there is no difference on effectiveness between title indexing and abstract indexing, then we can suggest using title of document as index rather

than using abstract. Although the storage getting cheaper, we still need more storage because of the growth of document in the internet is very fast [6].

We can evaluate the effectiveness of indexing based on retrieved document. For example, the retrieved documents for the SI query [nuklir unsur kelumit air], the query in [12], that based on Boolean technique on abstract indexing can be seen in the Table 1.

Table 1 showed that title of document contained terms in the query, or in other way there was similarity terms on query and title of retrieved documents. Document 0251 and 0259 are document that judged relevant contained terms on query. The title of document 0135 that judged not relevant does not contain terms query.

Based on that, intuitively, with title indexing, the relevant document can be retrieved. So we can expect that the effectiveness of title indexing and abstract indexing in terms of producing relevant document have no difference. But to what extend that the effectiveness would be achieved? To answer this question, we did this experiment.

The rest of the paper is organized as follow. We present and discuss some

background and related work on indexing in the next section. The methodology is presented in section 3 in which we detail the process we used. Result of our experiment and analysis we described in section 4. Finally we conclude our research in the section 5.

Table 1. Retrieved document from SI query from Mustangimah's experiment

Docu-ments ID	Title of documents	Relevance
0135	Metode Spektrometri Pendar Sinar X untuk Penentuan Zr dari Hasil Olah Pasir Sirkon Setelah Diendapkan sebagai Zr-Kupferonat	NR
0251	Studi Perbandingan Kandungan Unsur-unsur Kelumit di Daerah Penderita Gondok Endemik dan Kontrol di Kabupaten Magelang dengan Metode Graphite Furnace Atomic Absorption Spectrophotometry	RM
0259	Penentuan Unsur Kelumit dalam Tanah di Daerah endemik Gondok Secara analisis Pengaktifan Neutron	R

Note. NR = Not Relevant, RM = Relevant Marginal, R = relevant

2. BACKGROUNDS AND RELATED WORK

Basically, indexing is the process to find terms that represents the document. Meulen defined indexing as process to transform document into numbers of terms as identifier to represent the content of document [11]. The retrieval process later depend on that terms, the chosen terms should be represent the content of document. If the chosen terms didn't enough to represent the document, it would decrease effectiveness of the systems. Meulen classified indexing into manual indexing and automatic indexing [11]. Manual indexing is indexing manually by human, known as indexer expert. Automatic indexing is indexing that done by computer-based system.

Indexer read document and pick some terms to be index based his understanding to the document. Indexing by indexer usually aims to see subject or "aboutness" of the document [10]. To get the good terms as index,

indexer has to read the entire document, but its not practical and time consuming. The other way, the indexer should read parts of important document and miss the rest (parts read and parts skim). Part of document include title, abstract, table of contents, introduction, open and close paragraph, conclusion, diagram, etc [10].

The manual indexing will cause inter-indexer and intra-indexer inconsistent. Inter-indexer happen when the same document will be judged different by another indexer because of their understanding to the document is different. Another weakness of manual indexing is intra-indexer, while an indexer chosen different terms to index a document in another time [13]. Intra indexer also caused by the richness of language, a concept can be express in many ways, for examples the occurrence of synonym, homonym, related terms, and thesaurus.

To avoid the inconsistency in manual indexing and to make indexing process faster, we used computer system to get terms as index from the document (automatic indexing). Likely manual, automatic indexing also concern on part of document to be indexed, more terms in the part document to be indexed, more terms index could be resulted. Parsing, stoplist, stemming and term weighting are the step in the automatic indexing [14].

In Hasibuan experiment [7], there are three components of document to be indexed and stored in the system. The components are title, abstract and subject of document. Title and abstract usually given by author, which contained author understanding about document (aboutness). Subject indexing given by indexer expert that read the document.

Pao claimed in the Hasibuan that effectiveness of indexing could be seen from relevance of terms index with document [7]. Katzer had compare effectiveness of indexing from title, abstract and combined of title and abstract [8]. He showed that there was little overlapping among retrieved documents from title, abstract and combines title-abstract indexing.

This research will compared title indexing and abstract indexing based on relevance of retrieved documents.

3. METHODOLOGY

In this section, we discuss method that we used in the experiment, process flow, network architecture and relevance judgment that we used.

3.1. Process Flow

The process flow of system depicted in figure 1. First, we have to build the network of index of document that describe in the section 3.2. The input of the system could be title or abstract of document and query from user. After the network is ready, system ask user to input the query according to format below:

*<weight>terms-1 <weight> terms-2
 ...<weight> terms-i ...<weight> terms-n*

Where weight between 0 and 1, and terms-*i* is the query terms. The value of query terms in vector query according to weight and the value other terms are 0. With this query format, user can be adaptive to give weight for each query term according to the significance of terms in query.

After that, query and index will be match. Matching process used activation method where every term in query will activate documents that relate to that term. After all documents retrieved then system will rank the document to get top 20 documents as output of the system. This output document number also used in previous experiments [3, 5].

3.2. Network Architecture

Network architecture used in this experiment similar to Kwok's Model network architecture [9], but different in number of layers. Unlike Kwok's architecture that uses three layers, this study uses just two layers in its network architecture. Query layer and term layer are combined become one layer.

Basically, every query given to the system is in the form of term(s). Figure 2 shows the network architecture that used in this experiment. It shows the relationship between term and document or described as term-document relationship. Every node in query layer/term layer contains term that has been used in documents collection. Every node in document layer contains document that constitutes document collection.

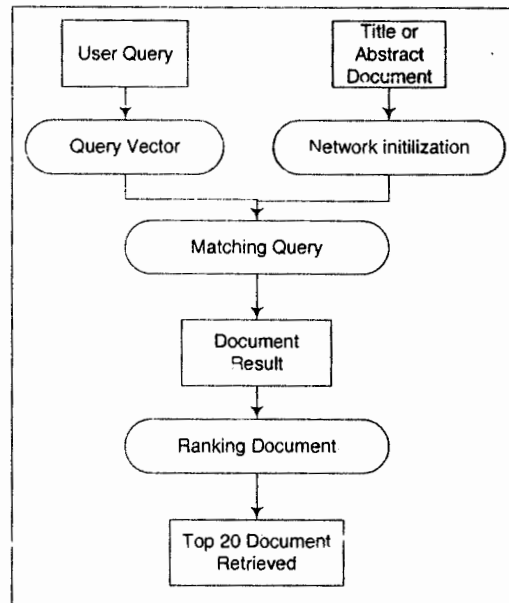


Figure 1. System Process

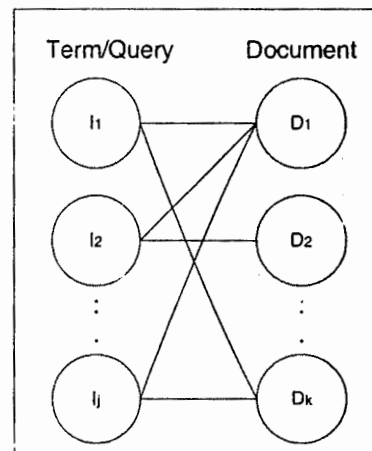


Figure 2. term-document network architecture

The edge that connecting between nodes in the query/term layer and nodes in the document layer, described as the weight of term-document relation. The weight of term-document relation calculated by using Savoy's rule [14], that also used in Azurat's experiment [3] and Andri's experiment [1]. Rule to calculate the weight of that edge is as follow:

$$W_{ik} = ntf_{ik} * nidf_k$$

where :

- ntf_{ik} = $tf_{ik} / \text{Max}_j (tf_{ij})$
- $nidf_k$ = $\log [n / df_k] / \log (n)$
- W_{ik} = weight term *k* in document *i*
- tf_{ik} = frequency of term *k* in document *i*
- n* = sum of document in all document collection
- df_k = sum of document that contains term *k*

$\text{Max}_j(\text{tf}_{ij})$ = frequency of the largest sum of term in a document

3.3. Relevance Judgment

This experiment used relevance judgment from Mustangimah's experiment [12], which used real user judgment for evaluating the retrieved document. The real user came from researcher was doing research in nuclear science and technology in BATAN. For each retrieved document, the users have been asked to judge the relevance of the document to the query. Three categories of relevance were used, which were "relevant" (R), "relevant marginal" (RM) and "not relevant" (TR). Table 2 describes definition of each category.

Added to that, this experiment also used subjective relevance judgment from Aribawono's experiment [2]. He used three categories relevance above and he was judged the relevance himself (relevance subjective).

Table 2. Relevance's Definition and Interpretation

Category	Definition	Interpretation
Relevant	Document directly related to the subject or research problem.	I very disappointed if the system can not find these documents
Relevant Marginal	Document related to the subject or research problem, but not directly.	I am not happy if system can or can not find these documents
Not Relevant	Document not related to the subject or research problem.	I disappointed if system find these documents.

4. EXPERIMENTS

We used documents collection from BATAN (Badan Tenaga Atom Nasional – National Atomic Energy Agency) as our corpus. Those documents collection consists of queries and documents written in Indonesian language. The corpus consists of 1162 abstracts documents with 2062 terms in title indexing and 9254 terms in abstract indexing. We used queries from study performed by Mustangimah [12] that listed in table 3.

We have been carried testing on both title indexing and abstract indexing. The results of experiment illustrated in table 4 to table 6. The activated document are all document resulted

for each queries and the retrieved documents are top 20 activated documents. We only judge the relevance of the top 20 activated documents (retrieved documents).

Table 3. List of Queries

Query Code	Query
S1	Nuklir unsur kelumit air
S2	Kristal tunggal silikon
S3	Lapisan tipis paduan logam Ti-Ni
S4	Lapisan tipis superkonduktor
S5	Pemungutan uranium pelat elemen bakar elektronis
S6	Tomografi radiografi neutron
S7	Limbah industri analisa aktivasi neutron instrumental
S8	Pengerasan permukaan bahan laser
S9	Penilaian sistem pengendalian reaktor daya
S10	Pemisahan isotop uranium

We could see from table 4 that 50% of document retrieved by both indexing. Title indexing resulted more document relevant than abstract indexing.

Based on table 5 and table 6 below, shown that the number of activated documents by abstract indexing more than by title indexing. It makes sense because the number of index terms on abstract indexing more than on title indexing. The average of relevant retrieved document on both indexing not quite different, it has only 1% difference. Based on this fact, we could say that effectiveness title indexing comparable with abstract indexing.

Table 4. Output based on query S1

Title Indexing		Abstract Indexing	
Document ID	Relevance	Document ID	Relevance
0259	R	0896	
0489	S-R	1125	S-RM
1125	S-RM	0128	
0398		0183	
0251	RM	0634	S-TR
1019	S-RM	0004	
0182		0385	S-TR
0487	S-RM	0461	
0184		0488	
0385	S-TR	0562	S-TR
0004		0889	S-TR
0488		0920	S-TR
0562	S-TR	1017	S-TR
0588		1134	
0595		1019	S-RM
0634	S-TR	0276	
0890	S-TR	1104	S-R
0896		0340	
1017	S-TR	0890	S-TR
1076		1027	

Table 5. Output based on title indexing

Query	Number of activated documents	Percentage	Number of relevant retrieved document	Percentage
		(%)		(%)
S1	116	9,98	6	30
S2	17	1,46	2	10
S3	46	3,96	1	5
S4	18	1,55	5	25
S5	141	12,13	2	10
S6	63	5,42	4	20
S7	114	9,81	2	10
S8	81	6,97	1	5
S9	204	17,56	5	25
S10	102	8,78	0	0
Average	90,2	7,76	2,8	14

Number of the same retrieved document on both indexing relatively high, 49.5% on average (see table 7). On S1 query, document with ID 0259, which judged relevant, retrieved on first ranking by title indexing but not retrieved by abstract indexing. The same situation occurred to document with ID 0251, which is judged relevant marginal, retrieved by title indexing but not retrieved by abstract indexing (see table 3). This fact showed us again that title indexing can be comparable with abstract indexing, even sometimes better.

Table 6. Output based on abstract indexing

Query	Number of activated documents	Percentage	Number of relevant retrieved document	Percentage
		(%)		(%)
S1	340	29,26	3	15
S2	71	6,11	1	5
S3	114	9,81	3	15
S4	51	4,39	4	20
S5	242	20,83	1	5
S6	113	9,72	5	25
S7	235	20,22	3	15
S8	264	22,72	2	10
S9	455	39,16	8	40
S10	161	13,86	0	0
Average	204,6	17,61	3	15

Based on this experiment, we may can conclude that title indexing can perform the same level effectiveness to abstract indexing although with lower number of index terms.

5. CONCLUSIONS

This experiment showed us several facts, which are:

- The numbers of index terms on title indexing 4.4 times lower than abstract indexing.
- The number of activated document on abstract indexing higher than title indexing (17.61%: 7.76%) (see table 4 & 5).

- The average relevant retrieved document on both indexing not quite different, 14% on title indexing and 15% abstract indexing.

Based on these facts, we can say that title indexing may achieved the same level effectiveness to the abstract indexing, so we can suggest that title indexing can be used to lower the storage index but with the comparable effectiveness.

Table 7. Intersection between title and abstract indexing

Query	Number of same document	Percentage (%)
S1	10	50
S2	9	45
S3	12	60
S4	12	60
S5	11	55
S6	10	50
S7	7	35
S8	11	55
S9	10	50
S10	7	35
Average	9,9	49,5

6. REFERENCES

- [1] Andri, Yofi, *Teknik Learning Scheme Berdasarkan Model P-Norm pada Sistem Temu-kembali Informasi*, Technical Report, Computer Science Faculty, University of Indonesia, 1997.
- [2] Aribawono, Anung, B., *Pendekatan Multi-dimensi Dokumen dalam Sistem Temu-kembali Informasi Menggunakan Model Spreading Activation*, Master Thesis. Computer Science Faculty, University of Indonesia, 2001.
- [3] Azurat, Ade, *Implementasi Sistem Temu Kembali Informasi dengan Melakukan Ekspansi Query melalui Inference Network Berbasis Probabilitas*, Technical Report, Computer Science Faculty, University of Indonesia, 1999.
- [4] Borko, Harold, *Toward A Theory of Indexing*, Information Processing and Management Vo. 13 pp. 355-365. Pergamon Press, 1977.
- [5] Budi, Indra, *Ekspansi Query menggunakan Constraint Spreading Activation dalam Sistem Temu Kembali Informasi*, Technical Report. Computer

- Science Faculty, University of Indonesia, 2000.
- [6] Denenberg, Ray, *Structuring and Indexing the Internet*, accessed from <http://www.loc.gov/z3950/agency/papers/italy.html> on December 10th 2002.
- [7] Hasibuan, Zainal A., *Document Similarity and Structure: Using Bibliometrics Methods and Index Term as Approach to Improving Information Retrieval Performance*, Dissertation, Indiana:Indiana University, 1995.
- [8] Katzer, J, et al., *A Study of The Overlap Among Document Representations*, Information Technology: Research and Development 2 (261-274); 1982.
- [9] Kwok, K.L., *A Neural Network for Probabilistic Information Retrieval*, Dept. of Mathematics and Computer Science, Western Connecticut State University, Danbury, CT 06810, 1989.
- [10] Lancaster, F.W., *Indexing and Abstracting in Theory and Practice*, Library Association Publishing. London, 1998.
- [11] Meulen, W. A. Van Der and P.J.F.C. Janssen, *Automatic versus Manual Indexing*, Information Processing & Management Vol. 13 pp. 13-21, 1977.
- [12] Mustangimah, *Efektivitas Sistem Temu-Kembali Informasi dan Analisa Bibliometrik: Aplikasi pada Dokumen Bidang Nuklir Berbahasa Indonesia*. Master Thesis. Post Graduates Faculty, University of Indonesia, 1998.
- [13] Rolling, L., *Indexing Consistency, Quality and Efficiency*, Information Processing & Management 17, no. 2: 69-76, 1981
- [14] Salton, Gerard, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, New York, 1989.