The Use of Inverted Index to Information Retrieval: ADD Intelligent in Aviation Case Study

Sodel Vázquez-Reyes, María de León-Sigg, Perla Velasco-Elizondo, Juan Villa-Cisneros, Sandra Briceño-Muro,

Autonomous University of Zacatecas, Software Engineering
Zacatecas, Mexico
vazquezs@uaz.edu.mx, mleonsigg@uaz.edu.mx, pvelasco@uaz.edu.mx, jlvilla@uaz.edu.mx, sgbm0592@gmail.com

Abstract. Nowadays store, index and retrieve information from document collections is a complex but necessary task. For this reason, information retrieval is fundamental to decision-making in companies. The *Be Intelligent* system offers a solution to storing, indexing and retrieval of documents content of ADD Intelligent Aviation company. The system performs searches through natural language expressions, presents the user a list of results containing document name, page, author, date and paragraph with search terms highlighted. The list of documents that meets the search is ordered by the relevance between the expression in natural language and the content of a document.
Be Intelligent system provides support for administration, indexing and retrieval of digital documents that the company uses during inspections of aircraft, reducing time to retrieve information.

Keywords: inverted index, information retrieval, precision, mean reciprocal rank.

1 Introduction

Presently is increasingly evident the urgency to search for information in large repositories that contain information in documents rather than data. This has transformed information retrieval into an important field of research, and in the development of computer systems that must a) be able to process quickly large collections of documents; b) allow flexible search operations, and c) allow classification of recovered information [1]. In this regard, the final purpose of information retrieval systems is to offer mechanisms that allow companies to acquire, produce and transmit, at the lowest cost, data and information with the attributes of quality, precision, and validity, in order to be useful to decision-making [2]. However, it is important to specify that the main function required from a retrieval information system is not to return the information desired by the user, but to indicate which documents are potentially relevant to his need of information, because, in fact, a user of an information retrieval system is interested about some subject, and not in the specific data that satisfy a query. Information retrieval systems deal with text,

J. Mejia et al. (eds.), *Trends and Applications in Software Engineering*, Advances in Intelligent Systems and Computing 537, DOI 10.1007/978-3-319-48523-2_20

generally written in natural language, not well structured, and semantically ambiguous [3].

Consequently, retrieved documents are judged as useful or not useful, and usefulness is judged in terms of degrees of effectiveness, being the standard measure the utility of the retrieved document [4]. As utility is a complex criterion to measure, because it is mainly based on the judgment of someone (user or non user) [5], several metrics are commonly used. Examples of them are *Precision*, *Reciprocal Rank* and *Mean Reciprocal Rank*. *Precision* (1) is defined as the proportion of recovery relevant documents [6]. This metric evaluates the system ability to position first most of relevant documents and measures the percentage of recovery documents that have relevance [6]. Its calculation is obtained with

$$Precision = \frac{number\ of\ relevants\ recovered}{number\ of\ results} \quad \textbf{(1)}$$

On the other hand, *Reciprocal Rank – RR* (2) is used to measure the system ability to retrieve relevant documents in the higher positions in the list of result. It is calculated by the next equation [7]:

$$RR = \frac{1}{rank(i)} \quad \textbf{(2)}$$

where *rank(i)* refers to the position of the document that contains the correct information to query *i*, and *RR* will be *cero* if information searched is not located in any document.

Finally, the *Mean Reciprocal Rank –MRR* (3) is the average or the *RR* values for all of the queries [7]. This metric gives the highest score to documents that are in the first positions of the list of results, because it measures precision and the order of the correct results [7]. This metric is calculated with

$$MRR = \frac{\sum_{i=1}^{|Q|} RR}{|Q|} \quad \textbf{(3)}$$

where *RR* is the *Reciprocal Rank*, shown in (2), and *Q* is the number of results. To facilitate retrieval of documents or parts of documents, indexes are created. Indexes, then, provide users with effective and systematic means for locating documentary units relevant to information needs or requests [8]. There are several index structures, but one of the most popular is inverted index. An inverted index is a mechanism oriented to words formed by two elements: 1) the vocabulary, defined as the set of different terms (words) in texts; and 2) the occurrence lists, defined as the list of documents in which a given term appears [1].

In this document is presented the use of *Be Intelligent*, a retrieval information system for company (*ADD Intelligence in Aviation*). The system uses an inverted index to retrieve information from a document collection stored in an external hard drive. This research first outlines *ADD Intelligence in Aviation* needs of information, describing briefly the use they give to it. This company represents the case study presented in this research. Next follows the description of the use of the inverted index data structure to retrieval information. Later the results of the implementation of the inverted index are presented. The document concludes by discussing the results of the implementation and the conclusions reached.

2 ADD Intelligence in Aviation case study

ADD Intelligence in Aviation is a center of aeronautical engineering specialized in physical inspections on airplanes. Nowadays, *ADD Intelligence Aviation* works with clients distributed in Mexican cities as Toluca, Ciudad de México and Sonora, but also develops projects in Seattle, WA., Santiago de Chile, in Chile, and Quetta, in Pakistan.

When an inspection project begins, engineers of *ADD Intelligence in Aviation*, gather different manuals, inspection templates, and other needed documents to verify an assigned aircraft. These documents are stored in different formats, including **.pdf, *.docx, and *.xls*. The document collection is fed with manuals developed by *ADD Intelligence in Aviation*, inspection templates, and airworthiness directives obtained in web sites of different aviation agencies, such as the International Air Transport Association, the European Aviation Safety Agency, and the American Airlines Maintenance. The document collection is stored in an external hard drive belonging the company. This was a company's internal decision. Access to this hard drive is made through *ADD Intelligence in Aviation* local networks. This situation implies that if an engineer is working outside the reach of local network, a collection of documents is send to him via e-mail, or by using a file-sharing service as Google Drive or Dropbox. Because all of this, it is needed a system to consult the content of documents and confirm if it has been useful to aircraft inspection.

3 Inverted index data structure to information retrieval

It is a well-known fact that building indexes is needed to implement efficient searches. A data structure called inverted index which given a term provides access to the list of documents that contain the term. The inverted index is the list of words and the documents in which they appear.

Most operational information retrieval systems are based on the inverted index data structure. This enables fast access to a list of documents that contain a term along with other information (for example, the weight of the term in each document, the relative position of the term in each document, etc.). In information retrieval, objects to be retrieved are generically called "documents". Given a user query, the retrieval engine uses the inverted index to score documents that contain the query terms. Terms that are considered non-informative, like function words (the, in, of, a, etc.), called stop-words, are often ignored.

Inverted index exploits the fact that given a user query, most information retrieval systems are only interested in scoring a small number of documents that contain some query term. Since all documents are indexed by the terms they contain, the process of generating, building, and storing document representations is called indexing and the resulting inverted files are called the inverted index. Building an inverted index for maintaining any kind of searching system requires to perform a series of steps: fetching the documents, removing the stop words, and finally merge and store the terms to inverted index [1]. This process of index construction or indexing is shown in Fig. 1.

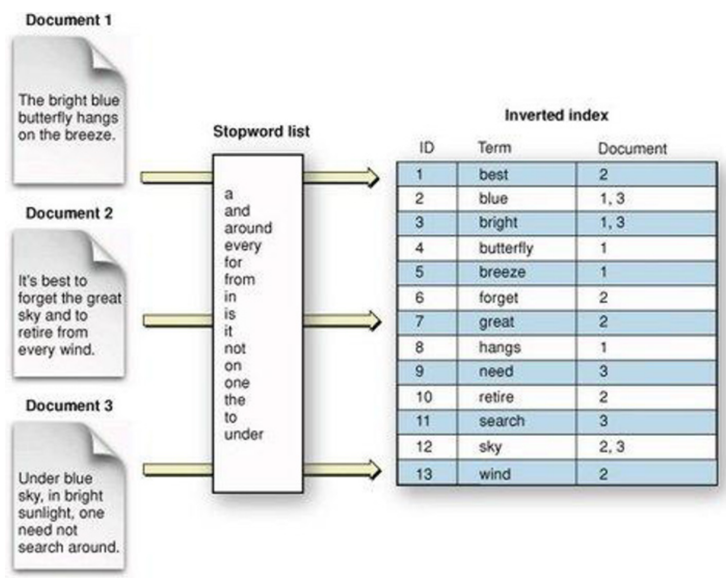

Fig. 1 Process of index construction [9]

4 Experiment design

As discussed in Section 1, there are several metrics to evaluate an information retrieval system. In this section are presented the values obtained for those metrics with the *Be Intelligent* system.

To evaluate *Be Intelligent* system, a collection of several documents was formed. The collection contained 300 documents: 100 of them contained information related to aircraft control, maintenance, and inspections; 115 documents contained information about software engineering, and 85 are documents with guides to research development and information retrieval. Also, two groups of users were formed, each one with five test users, as described next:

Group 1: Students of software engineering, inexperienced in information retrieval systems, without a specific need of information.

Group 2: Authentic engineers of *ADD Intelligence in Aviation*, with specific needs for information to do assigned inspections.

Tests where organized in two phases. Phase one corresponds to a short query; phase two corresponds to long queries. This organization allows evaluation of relevant documents obtained by *Be Intelligent* system. Each phase consisted of the next steps: a) access *Discover* option in *Be Intelligent* system (a system screenshot is shown in Fig. 2); b) search for a concept or phrase depending on phase a system screenshot is shown in Fig. 2); c) review retrieved documents; d) record count of total documents found, count of total relevant documents identified, and position of most relevant document retrieved in a web form after each query.

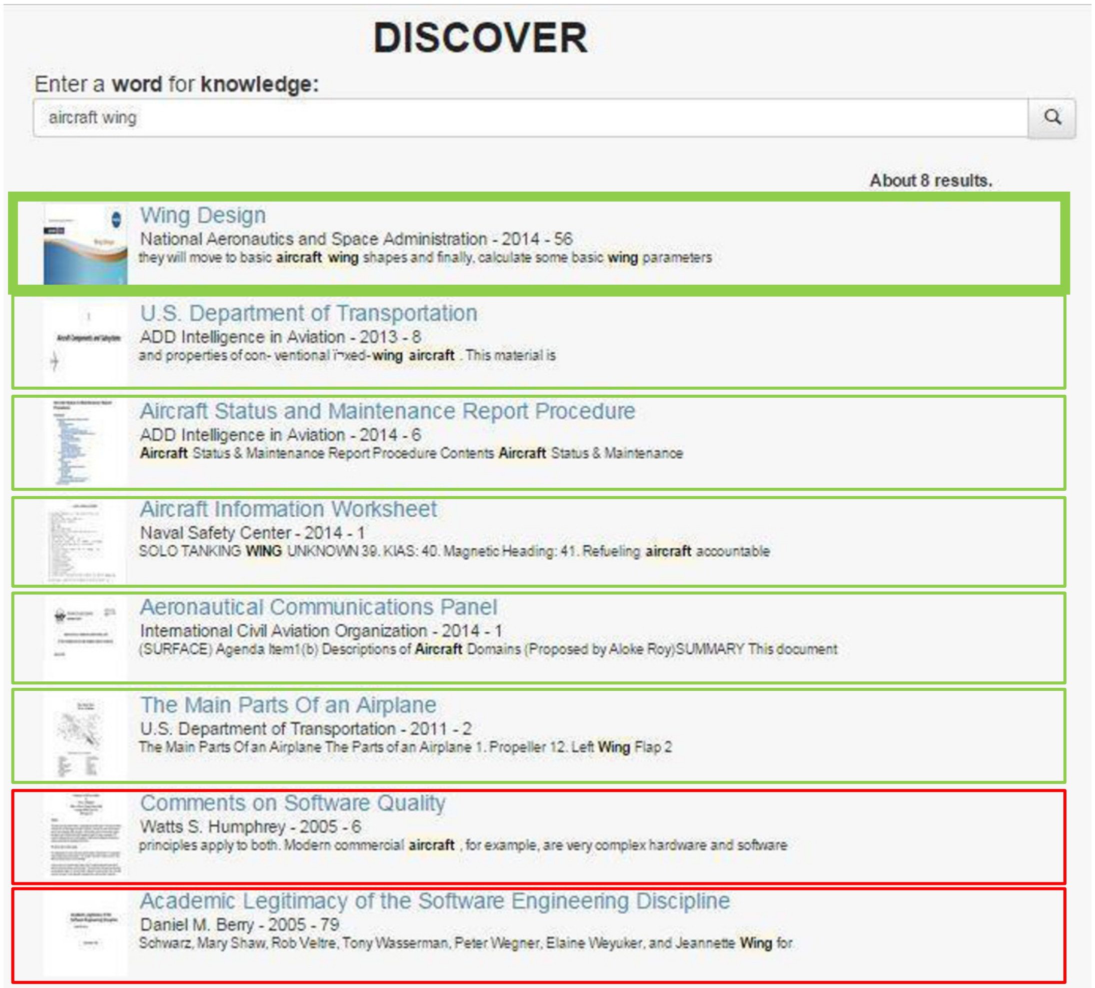

Fig. 2 Screenshot showing documents retrieved by Be Intelligent system when "aircraft wings" terms are searched

5 Results

Results obtained for each group and each phase (described in Section 4), are shown next.

5.1 Group 1 results

Phase 1. Fig. 3. shows calculated *Precision* to Group 1-Phase 1. No search obtained *1.0*, and average *Precision* was *0.73*. Best results were obtained by test users three and five, with a value of *0.78*. This value is considered respectable because the query domain of this group was wider.

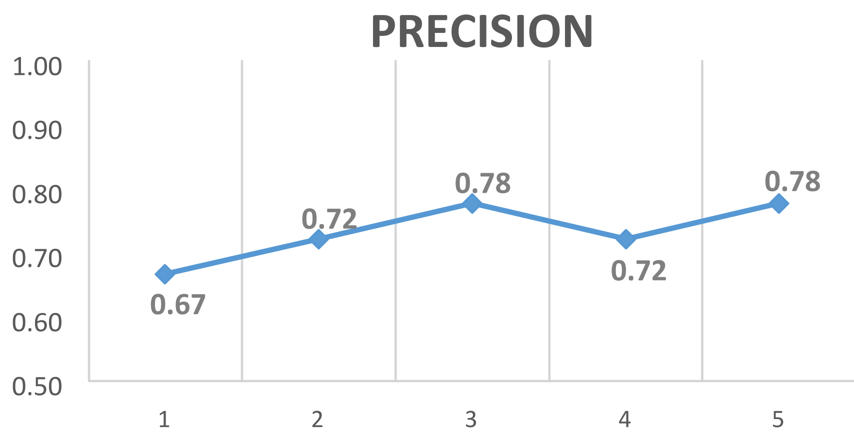

Fig. 3 Calculated Precision to Group 1, phase 1

Mean Reciprocal Rank calculated for queries of Group 1-Phase 1 is shown in **Table 1**. In this table is noted that first relevant document was located in third position, in queries three and four. With this data, *MRR* was *0.27*. This value was far from the ideal value of *1.0*, but it should be considered that this group experience level is low so its queries were less specific and the query domain was wider. Due to this, *Be Intelligent* responded adequately in this context.

Table 1. Calculated Mean Reciprocal Rank with results of Group 1-Phase 1

User	Rank	RR
1	4	1⁄4 = 0.25
2	5	1⁄5 = 0.20
3	3	1⁄3 = 0.33
4	3	1⁄3 = 0.33
5	4	1⁄4 = 0.25
	MRR	= 0.27

Phase 2. Results obtained with a long query done by Group 1-Phase 2, are shown in Fig. 4. In this figure can be observed that users one and two got *1.0* precision, while the average precision value is *0.89*, it is very close to the ideal value of *1.0*.

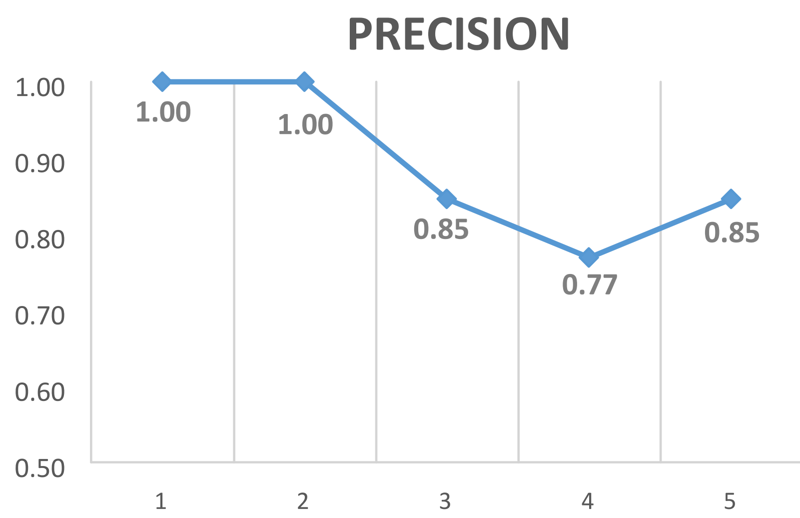

Fig. 4 Calculated Precision to Group 1-Phase 2

MRR for retrieved document in this phase was *0.29*, as shown in **Table 2**. Results with best-positioned documents were queries one, three, and five, where the first document was found in third position. *MRR* value is expected because the system does not respond to long queries with a group of a low level of experience.

Table 2. Calculated Mean Reciprocal Rank with results of Group 1-Phase 2

User	Rank	RR
1	3	1⁄3 = 0.33
2	5	1⁄5 = 0.20
3	3	1⁄3 = 0.33
4	4	1⁄4 = 0.25
5	3	1⁄3 = 0.33
	MRR	= 0.29

5.2 Group 2 results

In this section, are shown results obtained from experiment with Group 2. Authentic engineers formed this group, and due to this, their level of experience with search systems is high, as well as their clear information needs.

Phase 1. In this phase, the established query was directive FAA-2013-0695, a very specific term. Average Precision was *0.67*, and users three and four obtained *1.0* Precision value. However, users two and five obtained precision values under average because they needed more details about the directive, as shown in Fig. 5. Accordingly with this figure, a 67% of retrieved documents were relevant, and due to this, the system responded adequately during this part of an experiment.

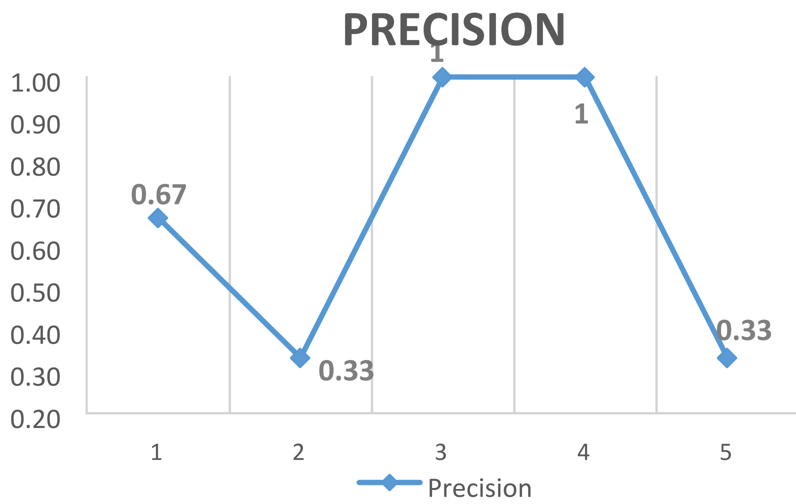

Fig. 5 Calculated precision to Group 2, phase 1

From data presented in **Table 3**, it is shown that *MRR* is *1.0*, because all test users indicated that first document shown was the most relevant for this query. The first retrieved document was judged as useful; for this reason, the system is judged effective.

Table 3. Calculated Mean Reciprocal Rank with results of Group 2-Phase 1

User	Rank	RR
1	1	1⁄1 = 1
2	1	1⁄1 = 1
3	1	1⁄1 = 1
4	1	1⁄1 = 1
5	1	1⁄1 = 1
	MRR	= 1.0

Phase 2. Data are shown in Fig. 6, the average precision for this phase of the experiment was *0.83*, indicating that more than 80% of retrieved documents are relevant, improving engineer decision-making.

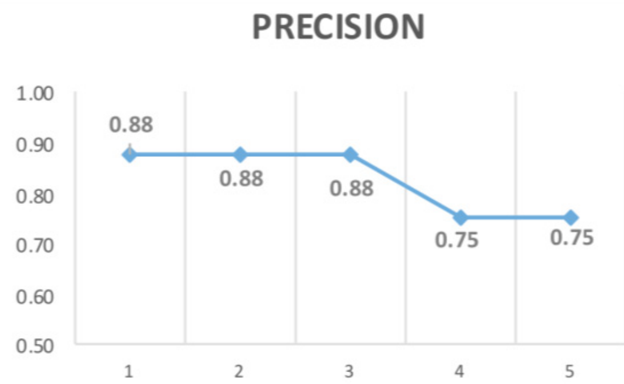

Fig. 6 Calculated precision to Group 2-Phase 2

In **Table 4**, are shown *RR* and *MRR* calculations. As can be observed, in 80% of queries document with most relevant information for the query is located in the first position. Due to this, the engineer found relevant information without loosing too much time reviewing other non-relevant documents.

Table 4. Calculated Mean Reciprocal Rank with results of Group 2-Phase 2

User	Rank	*RR*
1	1	1⁄1 = 1
2	1	1⁄1 = 1
3	2	1⁄2 = 0.5
4	1	1⁄1 = 1
5	2	1⁄2 = 0.5
	MRR	= 0.80

6 Conclusions

Computer systems to retrieve information from a large collection of documents must be able to reduce search time and allow classification of recovered information at the lowest cost, but with high quality, precision and validity to be useful to decision-making. *ADD Intelligence in Aviation* needed such a system, and in this research, results from the implementation of an inverted index data structure in the *Be Intelligent* system were shown. Two main contributions can be identified for *Be Intelligent* system: first, the creation of a storage and indexing scheme that facilitates information management of documents of routine use in a company. The indexing scheme supports workflow to include new documents in the indexing structure, the core of *Be Intelligent* system.

Second, an easy, quick and precise method to access their document collection was provided. Users now can retrieve information needed to achieve an aircraft inspection, without the extra steps required to do a search on an external hard drive and without loosing too much time reviewing non-relevant documents.

In this way, the application of an inverted index is helping a real company to achieve its duties in a more productive form.

We can describe one future direction to update the project, improvements of retrieval information with an algorithm for a passage retrieval approach. It could consist of four points: 1.- Execute a full document retrieval, 2.- Split the top retrieved documents into passages, 3.- Execute passage retrieval against the passage set created at point 2 and 4.- Return the top retrieved passages. To decide whether a passage retrieval strategy is useful or not, it is necessary to evaluate their ability to mine passages effectively.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[2] J. A. Arévalo, "Gestión de la Información, Gestión de Contenidos y Conocimiento," in *II Jornadas de Trabajo del Grupo SIOU*, 2007, pp. 1–15.

[3] P. Lara Navarra and J. A. Martínez Usero, *Agentes Inteligentes en la Búsqueda y Recuperación de Información*, Segunda Ed. Barcelona, España: Planeta-UOC, S. L., 2006.

[4] D. C. Blair, "The data-document distinction revisited," *ACM SIGMIS Database*, vol. 37, no. 1, pp. 77–96, 2006.

[5] R. Arquero Avilés and J. A. Salvador Oliván, "La Investigación en Recuperación de Información: Revisión de Tendencias Actuales y Críticas," *Cuad. Doc. Multimed.*, no. 15, pp. 2–3, 2004.

[6] F. J. Martínez Méndez, *Recuperación de información: Modelos, Sistemas y Evaluación*. Murcia, España: KIOSKO JMC, 2004.

[7] M. Levene, *An Introduction to Search Engines and Web Navigation*. Wiley Publishing, Inc., 2010.

[8] J. D. Anderson, "Guidelines for Indexes and Related Information Retrieval Devices," Bethesda, MD, 1997.

[9] C. Martín-Daucasa, "Desing and Evaluation of new XML Retrieval Methods and their Application to Parliamentary Documents," Universidad de Granada, 2012.