An Approach to the Design and Evaluation of an Enterprise Search Application

Daniel Zilio[1], Maristella Agosti[2], and Daniele Turato[3]**

[1] Department of Cultural Heritage, University of Padua, Italy
daniel.zilio@unipd.it
[2] Department of Information Engineering, University of Padua, Italy
maristella.agosti@unipd.it
[3] Department of Research and Development, SIAV, Rubano, Padua, Italy
daniele.turato@siav.it

Abstract. The paper reports on an experience of designing and implementing an enterprise search application. One of the motivations for the work is the lack of previous studies that deal with the problem of evaluating an enterprise search application. The reason for this lack depends on the building a significant test collection, because each company may be interested in dealing with different types of topic and data format possibly with atypical content formats.

Keywords: enterprise search, enterprise search application, document formats, enterprise search evaluation, test collection

1 Introduction and Motivations

Enterprise search is the term commonly used to define information retrieval in a business setting, offering users a way to search for informative content generated within their companies. By comparing it with *web search* and *desktop search*, we can identify some distinctive features: the content is stored in different information systems (e.g. data sources like customer relationship management software (CRM), enterprise resource planning software (ERP), intranets), different file formats are used (e.g. pdf, docx, xls), including both structured (e.g. database) and unstructured content (e.g. scanned documents), and different access levels give users different rights.

The purpose of enterprise search is to enable users to effectively find the information they need to perform their tasks, while at the same time requiring a minimal effort for the users and low costs to be sustained by the company in terms of inefficiencies [1,2]. Many of the problems of enterprise search have been addressed in previous experiences, as, for example, those reported in [5–7]. The notion of *relevance*, for example, can be different from that used in web search where there are usually many documents relevant to a query, and the ranking

** Siav S.p.A. has supported this work by providing the data and expertise in document management needed to accomplish the study.

tends to favor the most popular ones. Instead the typical enterprise search query has few correct answers. Dealing with enterprise information content has also some benefits to be exploited: the content is produced for dissemination purpose, unlike web content which is usually written to attract people. Moreover, we can obtain more contextual information on the queries: since the users are inside a company, we can obtain detailed profile information (e.g. role, experience, skills, team) or very precise location inside the company's building. Nonetheless, more effective algorithms that leverage those kind of benefits have to be developed.

One of the motivations for the work reported here is the lack of previous works that deal with the problem of evaluating an enterprise search application. The reason for this lack may depend on the difficulties in building a significant test collection: each company can be interested in dealing with very different types of topics and data format possibly with atypical content formats (e.g. in the case of the study reported in this paper CRM events were also dealt with), and the company may only be able to allocate limited resources to this type of activity.

In the present work we present a real case study of the design, development and evaluation of an application of enterprise search within a company, and draw attention to the issues that had to be overcome to complete the task. The paper also details the technological stack employed to build and test the enterprise search application.

In Figure 1 we sketch the different steps to design and implement the enterprise application together with its evaluation. The figure can also be used as a sort of outline of the paper.

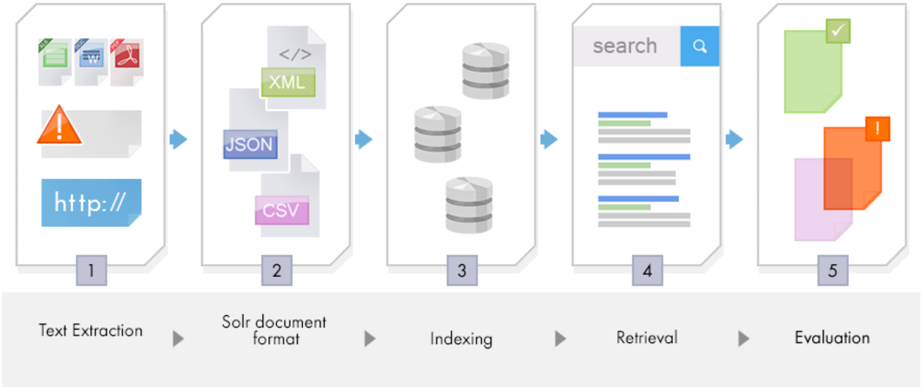

Fig. 1. Scheme of the steps to design, implement and evaluate the enterprise search application.

2 Related Works

The bibliographic research we conducted showed a lack of work on assessment and evaluation methodologies of *enterprise search*. This may be due to the fact that the importance of conducting an assessment to contribute to the improvement of the quality offered by information retrieval services has not yet been perceived, especially in the industrial sector. We must also consider that most of the research in this field is being carried out within companies, so we cannot exclude the possibility that many companies simply prefer not to disclose results and information considered to be strategic.

The results we present are based on [10] which reports on the study of designing and implementing a software prototype of an enterprise search application. The software prototype has been made interoperable with the Archiflow[4] document management application that is a software product of SIAV. The work has also been inspired by the example of using a real experimental collection to analyze the behavior of a system illustrated in [8], and by the work to design a model of a graph-based enterprise search engine provided in [9].

3 The Design of an Enterprise Search Application

The objective of this work was to produce a prototype able to perform the effective retrieval of documents that constitute the different collections present within the SIAV internal installation of the document management platform Archiflow. The system is designed to integrate the collections of documents with other content from sources other than Archiflow itself, in order to provide the end user with a spectrum of documents related to a specific subject of interest. An example which clarifies this aim is a search that requires the documents related to a particular customer: in addition to providing these documents, while maintaining efficacy as a metric, the system will provide additional related contents, like the latest tickets opened by the customer (available in the helpdesk management system), recently opened business opportunities with the customer derived from CRM or news on the web concerning the customer. This allows us to anticipate possible information needs, thus obtaining an advantage in terms of efficiency, in accordance with a business intelligence perspective.

The reference users are the professionals within SIAV. The company made available a set of data and documents from different information sources, including those managed by the different systems available within the company; in this way we were able to use these information sources to build a collection to evaluate the system.

The platform adopted is Apache Solr [3], which has been chosen after a benchmark process that saw Apache Solr emerge as a suitable tool for industrial applications.

[4] `https://www.siav.com/software-solutions/archiflow/`

3.1 The Different Data Sources and Their Formats

The different data sources that have been considered are:

- *Documents*: The documents are a subset of the document base of SIAV, that is, a part of the documents that are managed by the local installation of *Archiflow* at SIAV headquarters. The documents are of different types (such as invoices, emails, project sheets, etc.) and of different formats (including `DOC`, `DOCX`, `TXT`, `PDF`, `TIF`, `MSG`, `EML`, `RTF`, `XLS`, and `XML`). The metadata of each document are stored in an associated card from which the metadata can be exported and extracted in `CSV` format. For this category of data sources each document, together with the metadata related to it, is considered a basic unit of information.
- *Events*: SIAV is equipped with an issue tracking system that collects the tickets regarding the technical problems reported by their customers. The company manages the ticket elements through a CRM, which allows the coordination of the activities of the staff working at the helpdesk and the technical department. In this way all the actions performed by the staff involved are recorded in the system. Those actions are named *events*. It is possible to export the events from the CRM, and each event is enriched with a set of descriptive attributes. A total of 70,475 events were considered and studied; those events were collected over a period of about five years.
- *Web Pages*: During the work only a focused portion of the Web was examined; since the study on this portion of the Web is preliminary, the related results are not reported here.

It is important to note that objects within different sources may actually contain related information; e.g. a document can be related to a particular project and also a ticket item can be linked to it. One of the requirements of the system was to allow the user to exploit this information. For all considered data sources, the reference text language is Italian.

3.2 Text extraction

Since the documents were in different formats, an ingestion utility had to be developed and used to generate a new version of each document, in a format usable by Solr, for the necessary preprocessing and indexing procedures.

The first part of the ingestion work was the extraction of the textual contents from the documents: for each document file the text was extracted and saved in a simple text file, so as to allow the subsequent loading and verification through Solr of the quality of the extraction procedures. This processing step has been supported by different technical tools.

The documents of the used collection were:

- communications (11,974)
- bids and tenders (59,763)
- project reports (4,962)

for a total of 76,699 indexed documents.

The permanent memory space necessary for the storage of the preprocessed documents is 2.13 GB while the permanent memory used to store the original documents was more than 60 GB. For each original document, the final document with a valid format to be imported into Solr was created by adding the metadata of the original document to its extracted textual content. To manage and index all the documents (textual and metadata documents) through the search engine the library package `SolrNet`[5] has been used.

Information on document number and permanent memory size needed to store the indexed documents are reported in Table 1 and 2.

Document Sources	
Total Archiflow Documents	76,699
Total Events	70,475
Total Document Number	147,174

Table 1. Indexed document sources.

Permanent Memory	
Archiflow Documents Size	2.13 GB
Events Size	0.28 GB
Total Memory Size	2.41 GB

Table 2. Size for indexed sources.

4 Evaluation

Evaluation in IR is a well-established practice that has been studied for many years [4]. In accordance with this practice, an IR system or an IR application is evaluated using a test collection C constructed specifically for a given task. Therefore, it is necessary to select a set D of documents relevant to the task, a set of topics T that reflect the real information needs of users in relation to the task and define the relevance judgements RJ, so that each document is considered relevant or not relevant to a given topic.

The test collection to be used for the evaluation has to be built before starting the evaluation experiments and it is composed of the triple: $C = D, T, RJ$.

In this case the possibility of using an existing test collection was rejected because there is no test collection that is suitable to the type of tasks of interest. Therefore the test collection needed to be built. To build the test collection different alternatives were studied and the following operational strategy, exploiting a pooling methodology, was eventually decided. Of the three categories of documents – documents from Archiflow, events and web pages – it was decided to use documents and events that are peculiar to the company; web pages were not considered strategic for the design purpose. With the support of a thorough analysis of the tasks of interest conducted with the company management, five topics of interest for each document category were defined, and each topic was translated by a domain expert into a query. The pooling was done using Solr,

[5] `https://github.com/mausch/SolrNet`

Terrier[6] and two *Desktop Search* named *Copernic*[7] and *Windows Search*[8]. For each of these search engines, the following actions were conducted: indexing of the document base; execution of queries to produce the run; calculation of the number of documents to be judged on the basis of different depths of cut.

The achieved results [10] confirmed the choice of using Solr, as the results are comparable to those obtained with Terrier, while the two desktop search systems proved to be inefficient, as they returned not relevant results. It has to be noted, however, that while the obtained results querying the event base were relevant, those relating to documents originating from Archiflow were not, and this could be the consequence of having defined only five topics, a number that has to be considered insufficient for this documentary base.

The definition of the document base to be used and the subsequent evaluation of the results raised some issues: the formulation of the topics was really challenging; the process of defining the queries required the systematic cooperation between domain user experts and IR experts; the calculation and verification of all the calculated metrics was supported by MATTERS[9], a software toolkit written in Matlab for information retrieval evaluation; many activities requiring great time and effort were needed to obtain the pool.

5 Conclusions

Before this project SIAV used to rely on third party products to provide its customers with advanced search features, because the base search module in SIAV Archiflow provided only metadata based search.

The purpose of this work for SIAV was twofold: (1) develop a more flexible retrieval system and (2) acquire the necessary knowledge to put the enterprise search application into production, maintain it and in the future be able to enrich it with new desired features, in order to cope with new search needs of its customers. These goals were positively met, but there is still the need of an evaluation platform to continue improving the developed system. While enhancing the search system, the developer will in fact need a tool to check the effectiveness of new versions. The methodology presented in this work can be a valuable starting point for the development of an automated procedure.

Acknowledgements

We thank the anonymous referees for their fruitful comments and careful reading, which improved this article.

[6] `http://terrier.org/`

[7] `https://www.copernic.com/`

[8] `https://msdn.microsoft.com/en-us/library/windows/desktop/ff628790.aspx`

[9] `http://matters.dei.unipd.it/`

References

1. Feldman, S., Sherman, C.: The high cost of not finding information. IDC White Paper (April 2003)
2. Feldman, S., Sherman, C.: The information advantage: Information access in tomorrow's enterprise. IDC, Adapted from Hidden Costs of Information Work: A Progress Report by Susan Feldman, IDC 217936, and from Worldwide Search and Discovery Software 2009–2013 Forecast Update and 2008 Vendor Shares by Susan Feldman, IDC 219883 (October 2009)
3. Grainger, T., Potter, T.: Solr in Action. Manning Publications, USA (2014)
4. Harman, D.: Information Retrieval Evaluation (1st ed.). Morgan and Claypool Publishers (2011)
5. Hawking, D.: Challenges in enterprise search. In: Proceedings of the Australasian Database Conference ADC2004. pp. 15–26 (January 2004)
6. Hawking, D.: Enterprise search. In: Baeza-Yates, R., Ribeiro-Neto, B. (eds.) Modern Information Retrieval, 2nd Ed., pp. 641–684. Pearson Educational, UK (2010)
7. Mukherjee, R., Mao, J.: Enterprise search: Tough stuff. Queue 2(2), 36 (2004)
8. Rowlands, T., Hawking, D., Sankaranarayana, R.: Workload sampling for enterprise search evaluation. In: Proc. of the 30th Annual Int. ACM SIGIR Conf. on Research and Development in IR (SIGIR '07). pp. 887–888. ACM, New York, NY, USA (2007)
9. Tosato, D.: Exploiting ERP systems in enterprise search. In: Proc. of the 7th Italian Information Retrieval Workshop, Venice, Italy (2016)
10. Zilio, D.: Progettazione e realizzazione di un sistema di enterprise search. Master thesis in computer engineering, University of Padua, Italy (2016)