
Tag Cloud Control by Latent Semantic Analysis

submitted by
Angelina Velinska

supervised by
Prof. Dr. Ralf Möller
Dipl. Ing. Sylvia Melzer

Technical University of Hamburg-Harburg
Software Systems Institute (STS)

Dr. Michael Fritsch
CoreMedia AG
Hamburg

Contents

1	Introduction	3
1.1	Motivation and objective	4
1.2	Use case	5
1.3	Outline	5
2	Document Management System	6
2.1	1	6
3	Semantic Annotation	7
3.1	1	7
4	Latent Semantic Analysis	8
4.1	1	8
5	Implementation and evaluation of results	9
5.1	1	9
6	Conclusion and outlook	10
6.1	1	10

List of Figures

1.1	Online Document Management System	4
-----	---	---

Chapter 1

Introduction

During the last decade there have been constant optimizations in information retrieval effectiveness, making web search the preferred source of finding information. A substantial part of information retrieval deals with providing access to unstructured information in various domains. Information retrieval (IR) refers to finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [1]. Many people today use methods from the field of IR when they use a search engine online, or search through their emails. In this context "unstructured data" refers to data which does not have a clear structure. As an example, unstructured data is the opposite of the structured data stored in a relational database.

Information retrieval systems vary by the scale at which they operate. Three main types of IR systems can be distinguished - for web search, for personal IR, and systems performing domain-specific search. In web search IR systems, IR is done over billions of documents, stored on millions of distributed computers. Personal IR is performed at the level of consumer operating systems. And in the case of domain-specific or enterprise search, IR tasks are executed on collections of documents stored on centralized file systems, while search over the collection is provided by a handful of dedicated server machines.

IR technologies find wide application - in search engines, for browsing or filtering document collections, for further processing a set of retrieved documents. Before retrieval the documents are indexed, otherwise at each search, they would have to be scanned through for each query. Generally said, the index maps the words or terms back to the documents where they occur. An interesting technique for document indexing and retrieval, which will be applied in the current work, is called latent semantic analysis (LSA). It indexes the document collection by representing it as a reduced matrix of words and documents. LSA representation improves IR performance with respect to a basic problem of word-matching search - synonymy, or the case when more than one term describe the same concept.

While IR deals with retrieval of documents, other systems manage content. Content management includes a set of technologies and processes that support the creation, management and publication of content in any form or medium. Content may be documents, multi-media files, or any other file types that follow content lifecycle and require management. Content management systems (CMS) vary depending on their purpose and target environments - there are CMS for the web, for enterprise, for

mobile devices, as well as CMS for management of document collections, also called document management systems (DMS).

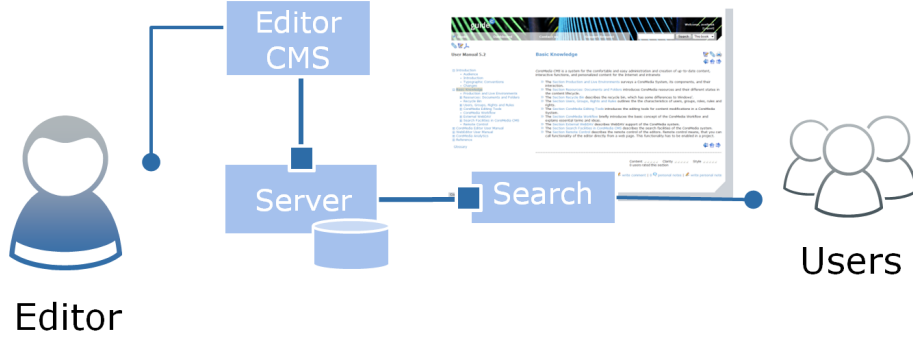


Figure 1.1: Online Document Management System

Figure 1.1 shows a simplified document management system, offering online access to its content. Using the editor CMS, editors can create, edit or delete content, which is managed and stored by a Content Management Server. Content is presented to the end users, who can access the document collection and search through it according to their information need. CMS usually consists of an environment for content management, and an environment for content delivery or presentation. However, in the introduction we only outline the basic concepts of the system. For a more detailed introduction, refer to Chapter 2.

1.1 Motivation and objective

A drawback of the classical LSA implementation as an information retrieval method is the low precision of the returned results. A previous work by David Mugo [2] has investigated the improvement of LSA precision performance by annotating the document collection and including the annotations used in LSA. In his work, Mr. Mugo constructs a concept-document matrix from the annotations used, and concatenates it with the word-document matrix normally generated in LSA process. The proposed solution, however, results in a slow speed of LSA, and has left Mr. Mugo's hypothesis open.

Taking into consideration the results from Mr. Mugo's work, the current project has several objectives to reach. It will investigate the implementation of LSA method for improving information retrieval in a domain-specific document management system with respect to context-based search. A further investigation will be made on improving the precision performance of LSA method by using semantic annotations, and on finding an adequate way to present the results of LSA as a tag cloud. And finally, it will be investigated how to use the tag cloud as a form of a relevance feedback to control LSA method.

In the context of the stated objectives, semantic annotations are meta data annotations used to add information to unstructured data, or to the document collection. Semantic annotations are based on an ontology in our case, specifically developed for the domain of interest CoreMedia CMS. Ontologies are used to capture some knowledge about a certain domain, by describing the concepts of the domain and the relationships between them. To further clarify the objectives, relevance feedback is an IR technique, used to influence the retrieved results based on the user's preference. It allows the user to modify the initial tag cloud by selecting the most relevant words. The tag cloud is then re-generated from LSA results with the relevance feedback posted as a query.

1.2 Use case

For the purpose of the current project, the investigations introduced in Chapter 1.1 will be performed in the environment of DocMachine 2.0¹, the document management system of CoreMedia AG², Hamburg. The system is based on CoreMedia CMS 2008, and contains manuals and guides which document CoreMedia CMS.

1.3 Outline

The reminder of this work is organized as follows. Chapter 2 describes in more detail what a document management system is, and provides an overview of the general structure of DocMachine 2.0, the DMS deployed at CoreMedia AG. Chapter 3 presents the basic concepts of ontologies and document annotations based on ontologies. In Chapter 4 an overview of latent semantic analysis method is given, as well as an approach for improving LSA's precision by including semantic annotations in the method. Chapter 5 presents the prototype implementation and makes an evaluation of the results achieved in this work. And finally, conclusions are drawn in Chapter 6, along with some limitations of the current study and outlook for a future research.

¹<https://documentation.coremedia.com/>

²<http://www.coremedia.com/>

Chapter 2

Document Management System

Summary. summarize me

2.1 1

Chapter 3

Semantic Annotation

Summary. summarize me

3.1 1

Chapter 4

Latent Semantic Analysis

Summary. summarize me

4.1 1

Chapter 5

Implementation and evaluation of results

Summary. *summarize me*

5.1 1

Chapter 6

Conclusion and outlook

Summary. *summarize me*

6.1 1

Bibliography

- [1] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.
- [2] D. M. Mugo, “Connecting people using latent semantic analysis for knowledge sharing,” master thesis, TU Hamburg-Harburg, Jan. 2010.
- [3] “Proceedings of the workshop on semantic search (semsearch 2008) at the 5th european semantic web conference (eswc 2008), tenerife, spain, june 2nd, 2008,” in *SemSearch* (S. Bloehdorn, M. Grobelnik, P. Mika, and D. T. Tran, eds.), vol. 334 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2008.
- [4] A. Micarelli and F. Sciarrone, “Anatomy and empirical evaluation of an adaptive web-based information filtering system,” *User Modeling and User-Adapted Interaction*, vol. 14, no. 2-3, pp. 159–200, 2004.
- [5] T. Landauer, P. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse processes*, vol. 25, pp. 259–284, 1998.