International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Content Based Document Information Retrieval System

Ajaykumar Ashok Awad

Department of Computer Engineering, PVPIT College of Engineering, Bavdhan, Pune, India

Abstract: ***The procedure with advancement of information surge has made it hard to get significant information on the web. In this proposed system, the necessity for practical Information Retrieval (IR) strategy has been extended. Document data contains huge information user can easily get the information by using only title and keywords of document or information. We propose a fast and effective content-based document information retrieval system that retrieves the information from the actual content of a document. In proposed system we use model of Latent Dirichlet Allocation that is used to extract major keywords for a given document. To improve the performance of system we use MongoDB database for the effective documents indexing. B-tree based indexing of MongoDB makes our system flexible, effective and fast than the previous system.***

Keywords: Information Retrieval, CBDIR, Inverted Indexing, B-tree Indexing, MongoDB

1. Introduction

Information Retrieval (IR) [1, 2] has been utilized as a part of numerous software engineering fields, and assumes a critical part on the web [3] inferable from the substantial number of information to recover. The customary web administrations are given IR. Clients can transfer archives or presentation documents alongside the title and portrayal. Likewise, clients can look the required data via seeking through the title or portrayal of the entered with IR [4]. Be that as it may, its ease of use and viability are restricted in that clients need to enter the correct catchphrase since it is looked just with the given title or portrayal. Besides, the title or description may not continuously contain enough data for the client to seek the essential substance. Keeping in mind the end goal to tackle this issue, it is expected to give the real substance data and additionally the title and depiction in IR frameworks. For this, in this paper, we propose a Fast and Effective based Document Information Retrieval framework (CBDIR). The fundamental points of interest of our framework are more adaptable furthermore, more successful and quicker recovery of data. In viewpoint of adaptability, our framework can without much of a stretch impart with generally utilized web stages utilizing the standard JSON design. We remove watchwords from the real substance of a report utilizing Latent Dirichlet Allocation (LDA) point display. Subsequently, the pursuit execution is enhanced contrasted with existing data recovery frameworks. Our recovery framework is very viable in that it utilizes real substance and also its title also, depiction. Our framework likewise gives clients required data continuously at a quicker recovery speed by utilizing upset ordering and B-tree based ordering.

2. Background

L. Azzopardi, M. Girolami, and C. Van Rijsbergen, "Topic basedlanguage models for ad hoc information retrieval," in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2004, pp. 3281-3286:

This paper explored the possibility of using a document specific term prior based on inferred topics induced from the corpus. The results show that on average the method was comparable to the standard language modelling techniques. However, when linearly combined with the background probabilities, in the two stage topic based Language Model, the IR performance was consistently superior to the standard models and standard two stage smoothing model across all collections. Due to the computational expense of the LDA method restricted this study to relatively small test collections. Further work is required to ascertain whether these results are consistent across larger and more varied test collections. Also, it is worth considering whether there is a significant difference between the variationalapproximations (LDA) or the maximum a posteriori estimate (PLSWLSA) for the aspect mod.[5]

X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 178-185:

Author proposed LDA-based document models for ad-hoc retrieval, and evaluated the method using several TREC collections. Based on the experimental results, we can make the following conclusions. Firstly, experiments performed in the language modeling framework, including combination with the relevance model, have demonstrated that the LDA-based document model consistently outperforms the cluster-based approach, and the performance of LBDM is close to the Relevance Model, which incorporates pseudo-feedback information. Secondly, we have shown that the estimation of the LDA model on IR tasks is feasible with suitable parameters based on the analysis of the algorithm complexity and empirical parameter selections. More importantly, unlike the Relevance Model, LDA estimation is done offline and only needs to be done once.[6]

O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," ACM Transactions on Information Systems (TOIS), vol. 29, p. 8, 2011:

Licensed Under Creative Commons Attribution CC BY

Author presented a novel approach to concept-based IR using ESA as a representation method, introducing a feature selection component that is based on pseudo relevance feedback. Paper have evaluated the proposed algorithms experimentally and demonstrated their improved performance. System also estimated the potential for further improving the results of this approach, and outlined several insights in this regard that can guide future work.[7]

3. Problem Statement

Information Retrieval is an essential step in the web information analysis. Content based information retrieval from document is achieve using the keyword based searching with indexing. Keywords are based on document user can find out easily the documents.

4. Proposed System

The proposed Content Based Document Information Retrieval System (CBDIR) is an information retrieval system that is based on the actual document contents uploaded by users. Here, a document represents any file in Portable Document Format (PDF), DOC, or PPT format. PDF is a file format developed by Adobe Systems, and DOC and PPT areMicrosoft MS Office file formats

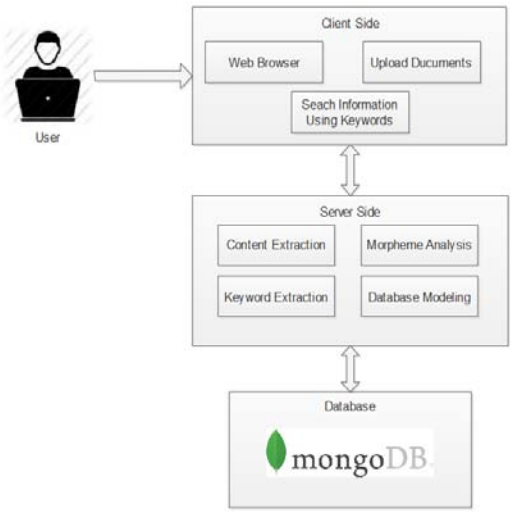

Figure 1: Proposed system architecture

The proposed system consist of following Methodology,

1)Connection
In this stage, a client communicates with the CBDIR system. A client sends all the information to the system such as the title, description, and document uploaded by a user, and the system responds to it. The connection is established through a pre-defined protocol such as TCP or HTTP. The data is transmitted with Java Script Object Notation (JSON) format.

2) Content Extraction
The raw contents in the transmitted document are then extracted. We focused on PDF-formatted files. To extract the content of PDF files, we used open source library PDFBOX which is provided by Apache.

3) Keyword Extraction
To extract meaningful keywords from the extracted nounsand verbs, we apply Latent Dirichlet Allocation (LDA) that isthe state-of-the-art topic-modeling algorithm. We used Malletopen source library for LDA [6]. Our system should process
one document promptly. LDA can extract topics from anindividual document in real-time. As the title and descriptionalready contain user-defined keywords, only the documentbody content is used for LDA topic modeling.

4) Morpheme Analysis
The separated crude content should be tokenized into morphemes. CBDIR tokenizes the info content into things and verbs. Other lexical classes are overlooked as a thing and a verb contains the most illustrative importance in the content. CBDIR utilizes openly accessible CoreNLP library as a morphological analyzer [11]. For preprocessing, we evacuated email locations, URLs, and exceptional characters first. At that point, CoreNLP tokenizes the sentence into lexical classes, furthermore thing or verb is changed over into lemma.

5) Database Modeling
We create database with extracted *P* keywords from LDA modeling. MongoDB based NoSQL is used as a database system. MongoDB has the virtue of flexibility, performance and scalability. MongoDB can store any type of data with key-document schema different from RDBMS. Thus,inverted indexing which is widely used in IR is adapted easily.

Scope:
1) Searching the document using content based searching approach.
2) Index the give documents using B-tree for fast retrieval.
3) Keyword extraction helps for indexing the documents.

5. Conclusion

In this paper, we proposed a quick and powerful Content Based Document Information Retrieval framework. We assessed the CBDIR framework with genuine information and researched its execution change over the pattern. In light of the test comes about, we demonstrated that our framework has three principle favorable circumstances. In the first place, our framework is effortlessly versatile to existing frameworks. It can without much of a stretch speak with a customers utilizing JSON organize. To bolster different sorts of information, we utilized MongoDB based NoSQL. Second, we extricated watchwords in record content successfully utilizing LDA point display and indicated upgrades in general execution. As there is most certainly not enough data in a title and portrayal, our framework moreover recovers the data of substance from archives. At last,by enhancing the quantity of catchphrases P that are separated from archive content, we effectively enhanced the recovery speed. Likewise we approved the effectiveness of our framework by contrasting it and the one without B-tree based ordering and the gauge that does not utilize any ordering blueprint. The outcome demonstrates that our framework is fundamentally superior to the current

Licensed Under Creative Commons Attribution CC BY

frameworks regarding both the general exhibitions and the recovery speed

6. Future Work

In our future work, other topic modelingtechniques such as Hierarchical Dirichlet Processes (HDP) and explicit semantic analysis [7] will be investigated forfurther improvement. We will also work on the improvementof LDA estimation speed for faster database construction.

References

[1] E. Greengrass, "Information retrieval: A survey," 2000.
[2] W. B. Croft, D. Metzler, and T. Strohman, Search engines: Information retrieval in practice: Addison-Wesley Reading, 2010.
[3] C. V. Forecast, "Cisco Visual Networking Index: Global Mobile data Traffic Forecast Update 2009-2014," Cisco Public information, February, vol. 9, 2010.
[4] C. Zhai and J. Lafferty, "Two-stage language models for information retrieval," in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 49-56.
[5] L. Azzopardi, M. Girolami, and C. Van Rijsbergen, "Topic basedlanguage models for ad hoc information retrieval," in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2004, pp. 3281-3286:
[6] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 178-185
[7] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," ACM Transactions on Information Systems (TOIS), vol. 29, p. 8, 2011:
[8] D. Comer, "Ubiquitous B-tree," ACM Computing Surveys (CSUR), vol. 11, pp. 121-137, 1979.
[9] C. Von der Wethand A. Datta, "Multiterm keyword search in NoSQL systems," Internet Computing, IEEE, vol. 16, pp. 34-42, 2012.
[10] Z. Wei-ping, L. Ming-Xin, and C. Huan, "Using MongoDB to implement textbook management system instead of MySQL," in Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on, 2011, pp. 303-305.
[11] K. Toutanova, D. Klein, and C. Manning, "Stanford Core NLP," ed: TheStanford Natural Language Processing Group. Available: http://nlp.stanford.edu/software/corenlp.shtml. Accessed, 2013.

Licensed Under Creative Commons Attribution CC BY