Journal of Medical Systems (2019) 43: 164
https://doi.org/10.1007/s10916-019-1294-5

TRANSACTIONAL PROCESSING SYSTEMS

Accurate Approach Towards Efficiency of Searching Agents in Digital Libraries Using Keywords

Vijayalakshmi Yellepeddi[1] · Manimegalai P[2] · Sasidhar Babu Suvanam[3]

Received: 23 January 2019 / Accepted: 15 April 2019 / Published online: 1 May 2019

Abstract
The age of information has done it simple for storing huge amount of data. In actual fact, a considerable segment of existing information is accumulated in the text databases that have huge set of documents from different sources like research articles, news articles, books, e-mail messages, web pages and digital libraries. In many text databases, stored data are in the semi-structured format in that they are neither entirely structured nor entirely unstructured. IR (Information Retrieval) field has been growing in parallel using database systems for several years. Contrasting to the databases system fields that have concentrated mainly on transaction and query processing of the structured data, IR is concerned with firm and retrieval of data from a huge quantity of text-oriented documents. Thus, IR tackles with unstructured and/or semi-structured databases. Information security requirements within a firm have experience major variations in the past some decades. By the establishment of computer, the necessary for automated equipment for securing files as well as other information that stored on the computer turned into evident. This is particularly in case of shared information resources via public network. This is the origin for having a secure computer system / the need for computer security. Computer Security can be achieved by Intrusion Detection Systems. In this paper, we address these issues by applying Similarity Search in two diversified fields: Digital Libraries and Computer Security. The paper discusses a fast and efficient similarity search technique for approximate retrieval of books metadata in Digital Libraries. In DLI the books retrieval takes place just by using metadata such as title, year, edition, author, publishing of a book. Though, if metadata is missing, incorrect or unfinished, then it creates the library retrieval system inefficient, incorrect leads too much confusion to the user. In this context even if the query from the user matches partially or fully with a stored pattern, the information related to that be retrieved. The paper talks about a method that functions rapid and effective, language independent, and flexible library retrieval system signature based similarity search. This system is able to retrieve not only the metadata that exactly matches the query but also fairly accurate identical because of missing words, jumbled words and spell mistakes. Fundamentally, signature file approach is used here. A signature file approach looks like the most capable for huge database as it has superior text retrieval features and requires little storage overhead.

Keywords Information retrieval · Information security · Similarity search

Introduction

The epoch of information has set it simple for storing huge amount of data. In actual fact, a considerable segment of existing information is accumulated in the text databases that have huge set of documents from different sources like research articles, news articles, books, e-mail messages, web pages and digital libraries. In many text

This article is part of the Topical Collection on *Transactional Processing Systems*

✉ Vijayalakshmi Yellepeddi
vijayasasi11@sngce.ac.in

Manimegalai P
manimegalai.p@kahedu.edu.in

Sasidhar Babu Suvanam
sasidharmails@sngce.ac.in

[1] Department of Computer Science & Engineering, Karpagam University, Coimbatore, India

[2] Department of Electronics and Communication Engineering, Karpagam University, Coimbatore, India

[3] Department of Computer Science & Engineering, SNGCE, Kadayiruppu, India

databases, stored data are in the semi-structured format in that they are neither entirely structured nor entirely unstructured.

Information retrieval and information security

Information Retrieval is defined as the knowledge of finding for information in manuscripts, finding for metadata that explains documents, finding for documents themselves or finding inside databases, whether hypertext network databases like WWW (World Wide Web) or Internet or separate data stores or intranets, for sound, data, images or text. There is general confusion; however, among document retrieval, data retrieval, text retrieval, information retrieval and every of these has its own theory, literature bodies, technologies and praxis. The most significant goal of IR is to retrieve related information depends on user demand on the external information source. IR is similar to most promising fields interdisciplinary, depends on mathematics, information science, computer science, linguistics, physics, statistics, cognitive psychology and library science. Automated IR systems are employed to decrease the overload of information. Many public libraries and universities employ IR systems in order to give access to journal, books as well as other documents. IR (Information Retrieval) field has been growing in parallel using database systems for several years. Contrasting to the databases system fields that have concentrated mainly on transaction and query processing of the structured data, IR is concerned with firm and retrieval of data from a huge quantity of text-oriented documents. IR is dedicated to searching for related documents and not searching for simple patterns or matches. Thus, IR tackles with unstructured and/or semi-structured databases.

An integrated technique for context-based and semantic retrieval

There are numerous search engines that built to deal with the problem of adopting traditional keyword-based search. A keyword based search technique employs user query to retrieve group of related documents from indexed document those suits the words provided by the user. Semantic Web is considered as an expansion of present web, where information gives well-defined sense that allows system as well as people for finer understanding and may allow to work efficiently through understanding information from various sources (Fig. 1).

In order to integrate various aspects of information needed for context-based and semantic retrieval, a representation model is constructed for cultural objects that broadly have three stages of information exposed in the figure:

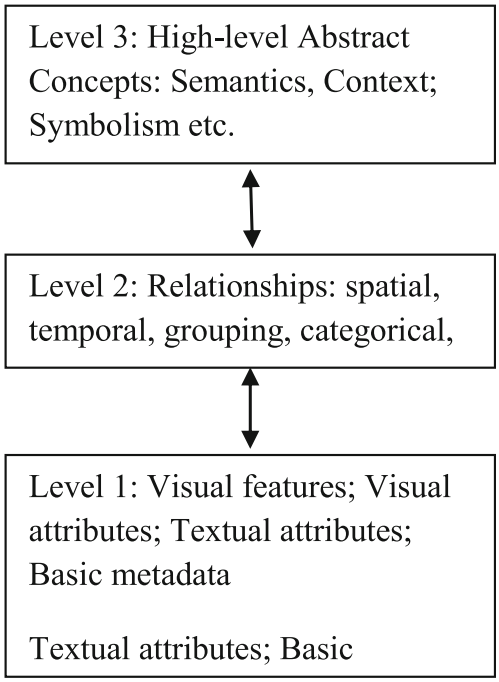

Fig. 1 Information levels for cultural objects

Applications of similarity search

In this thesis, we address these issues by applying similarity search in two diversified fields: Digital Library and Computer Security. Similarity Search is one of the effective information retrieval techniques that have turned out to be a basic computational task in different application regions such as multimedia information retrieval, digital library, pattern recognition, data mining, computer vision, data compression, statistical data analysis, biomedical databases and machine learning. In such kind of atmospheres, a precise match has small meaning and distance or proximity (similarity or dissimilarity) concepts are usually much more successful for searching. By the raising assortment of digital data kinds that practically covering all types of fact representation, automated data processing should value these natural guidelines and give sufficient tools for the similarity searching.

Digital libraries and its challenges

In this paper discussed the advantages and challenges of digital library information retrieval with the help of keywords.

Benefits of digital library are as follows:

- Limit less storage
- No physical margin
- Instantaneity of retrieval
- Multiple access
- Encircling the clock availability
- Indestructible

- Conservation and Protection
- Accurate value addition

In 2002, the DLI initiated book digitization with main intention of conserving the age-old and rate collection of books linked to Geography, Agriculture, Indian History, Technologies,, Medical sciences, Culture, Philosophy, epics between others, by digitization and prepared available online for research scholars, students and other common person at free of price. The digitization dreams the huge knowledge of human kind and creating it existing online has turned into an achievable goal, at present. But, there are some challenges in attaining this goal still. When 'the Million book project' was started, it was said to be the first of its type ever envisioned. The challenges include:

Incorrect and incomplete metadata

In DLI project, many books scanned are acquired from sources such as government archives and libraries and thus has metadata entered through knowledgeable people, who may be depends upon, but it is still arguable because of individual source. Because of these diverse book flow sources in DLI project in many languages and because of inadequate of standard formats, metadata is incorrect, missing or unfinished or sometimes complex to understand. Imprecise metadata obstructs successful search as well as retrieval of books, classification.

Manual errors

Manual errors are arising mainly because of miscommunication among human resources or because of their non-adherence or incompetence to practice and standards. These are the major and maybe the costliest of the issues, which might be viewed. For instance, librarians couldn't be well-advised regarding the ontology and hierarchy of book categorization and so could categorize book possessions to a type of 'Art' when it actually possesses to a type of 'Music'. The outcome would be that a book doesn't reveal up a fake search to a people who looking for 'Music'. An even bad condition might be allocating a wholly unrelated type.

Machine errors

Machine errors are creep in because of the comprehensible restrictions of the improper configurations or software of the software and machines. Errors because of improper configuration of scanners at the period of the scanning stage are more severe as the data created turns into less valuable because of low quality.

Indian languages

Various organizations employ diverse formats, like ITRANS, ISCII, Unicode, OmTrans, etc., to code the documents using Indian language. Exact search as well as rapid retrieval of digital content beside the user questions is a most important research problem.

Existing techniques

In [1], authors have proposed effective Focused Web-Crawling technique for search engine. The proposed technique traverses the net, opting for related pages to a predefined topic as well as ignoring those beyond the concern. Gathering domain-oriented documents by focused crawlers have obviously been assumed as the most significant strategies to determine appropriate information, whilst surfing the web is complex to manage irrelevant pages and to forecast that links bring about quality pages. The most focused crawlers employ local searching algorithm in order to traverse the net space, but they might be trapped easily within the limited subgraph of web, which surrounds the starting URLs as well. There is issue pertaining to related pages, which are ignore when no relations from starting URLs. In order to deal with this issue, they have proposed a focused crawler, in which measuring the rate of topic keyword and also estimate the synonyms with sub-synonyms of keywords. Here, the weight table was built in accordance with the user query.

Researchers, in [2] have developed federated search for text-oriented digital libraries in the hierarchal peer-to-peer systems. Peer-to-peer systems are potentially more powerful model to develop large scale networks for text-oriented digital libraries, but these systems have yet given very constrained support to text based federated search for digital libraries with the help of relevance based rating. This article has addressed the issues of resource rating and selection, resources representation and results combining for federated seek out for text-oriented digital libraries in the hierarchal peer-to-peer systems. Some of the existing techniques for text-oriented federated search were adopted and new techniques are proposed for resource selection and representation along with distinct features of hierarchal systems. The experimental findings of this paper has shown that the developed techniques has provided a better mixture of efficiency and precision than that of more general substitutes for federated search to text-oriented digital libraries in the peer-to-peer systems.

A research was carried out to develop a way out for suggesting digital library services depends on the data mining approaches like predictive classification and clustering techniques. Initially, similar users were grouped together depends on their profiles with search behaviors; and after that predictive classification for suggesting suitable services to those

users was employed. It has been revealed that users in similar group have a great probability of taking on same patters or services. Here, datasets were employed from KOBSON digital library. If the relation among the institution and the service were greater, then it would have enhanced precision. The findings have indicated that Naïve Bayes classification and k-means clustering can be employed to enhance the precision of service recommendation. On the whole precision is fulfilling, whereas average precision based on the particular service. The findings were enhanced for regularly arising services [3].

An analysis was conducted content-oriented online searches in huge database in multimedia documents like tables, currently text and color images using huge merged querying as well as retrieval system. Queries are presented as compound sentences in the natural language. They are converted into language of target database by domain-specific and world-specific knowledge. The documents recovered from database were examined for their information-content appropriate to user query in a semantic manner. There is no need of pre-indexing when new texts are stored in the system: if user query is necessary, then each search technique can be used to each single manuscripts that stored in the database. This holds a great range of flexibility concerning to the queries, which can be inquired. It involves that both vast quantities of data should be estimated in the short time period and that intellectual caching policies should be used. Hence, it is obligatory that the scheme prepared with high speed hardware processors conducting the most often used functions on both textual documents and images [4].

The authors, in [5] have disagreed regarding the significance of using expensive resources to generate local meta-data records, especially for academic materials, which have complete text search alternatives in the field of library science. The study has coordinated search question logs to complete text of items and the indexes of meta-data records. The findings have revealed the allocation of item detections, which were depends on the complete text solely, on the meta-data solely and on the mixture of both. Though, many search queries had extended beyond results in complete text and meta-data, some of the item detections happened by meta-data values alone or the mixture of complete text and meta-data. This has been suggested that generating local records does prop up item detection and retrieval.

In [6], researchers have proposed an incorporated semantic-based technique with reference to improve the retrieval efficiency in concept-oriented video retrieval. Multimedia content has been developing rapidly and video retrieval is considered as the most popular problems in multimedia study. To retrieve a required video, users convey their desires with respect to queries. Queries may be on motion, texture, audio, color, object and so on. Low level video representations are varied from greater level ideas where a client relates with video. Hence, query relied on semantics is considered as more tangible and realistic for end user. Understanding the query semantics has opened as a novel insight in bridging semantic gap and video retrieval. The proposed technique is relied on the incorporation of corpus based and knowledge based semantic word similarity measures to get back video shots for ideas whose marginal notes aren't accessible for system. Here, the dataset of 'TRECVID 2005' was employed for the estimation purpose and the findings of applying developed technique were compared with individual corpus based and knowledge based semantic word resemblance measures that were employed in the prior analyses in the similar domain. The authority of incorporated similarity technique is revealed and estimated with respect to MAP (Mean Average Precision).

A study was attempted to develop a Smart Digital Library that used by broader society everywhere they are. This system is developed in Smart Digital Library gateway model that employs semantic similarity technique to find out articles, journal or books through author name or titles. This technique is mainly employed to identify the suggested books to be read through viewers of library depends on the authentication from a prior reader routinely. There are several steps involved in digital library system development such as analysis, design, testing and implementation phases. In analysis stage, WebQual is used for preparation of devices that to be given to respondents and data acquired from respondents would process with the help of quality function deployment. The following were the findings that drawn from this analysis:

i. Data processing results have determined customer needs including necessary for digital library and component of customer needs contain feature information search through name, title and year by assuming the similarity of semantics or meaning, testimonials or room features opinions, books ratings depends on the testimony; and
ii. Depending upon the data processing results at technical requirements, an analysis of Quality Function Deployment has found that there was a number of design features, in which digital library isn't significant in the perspectives of some respondents. Digital design library is prepared to be constructed along with results by neglecting features that aren't needed by the user [7].

Researcher in [8] have discussed about the four most significant challenges for digital libraries include:

i. Interoperability: This has become progressively more significant as increasingly content moved online as well as demand for combined access grew. Interoperability has fall into three dimensions include acquiring disparate schemes to operate together really, allowing software to operate on various schemes and supporting transform of content among systems;

ii. Community engagement: In several ways, community engagement with opportunities that digital libraries have played stronger social tasks. The field of digital library has built huge technical progress and enhanced the accessibility and discoverability of academic as well as cultural heritage content. A vital challenge of this field is commitment with people they serve and raising digital libraries' values;
iii. Intellectual property rights: The legal structure for securing intellectual property rights has become a huge challenge for digital libraries. With a huge quantity of different content like images, text, video, audio and more, radical changes have taken place regarding how people and systems interact, generate and communicate, exchange, relate to content and reuse. This dynamic, high speed, participatory online information atmosphere attain benefits from open systems and simple, low barrier sharing and swap over of digital content.
iv. Sustainability:

The field of digital library knowledge is mainly about how to construct digital libraries outpaces its recognition of how to maintain them. While the developers of digital libraries desire to provide open access to all people, digital libraries aren't open for their developers to generate and sustain. These prices should be recovered in some way. Financial sustainability is very critical [8].

Similarity search – methods and techniques

An accurate match has some meaning, and distance or proximity (similarity or dissimilarity) concepts are usually more successful for searching. In traditional DBMS, the search gives accurate results, i.e., an item either belongs to result-set or it doesn't. Since the traditional partial, exact, and series retrieval paradigms are not succeed; so fulfill the content-based retrieval desires of many rising data processing applications and the concepts of proximity or similarity are turning into progressively more relevant.

Need for similarity search

Similarity Search is frequently employed for small user queries in search engines like Lycos, AltaVista, Yahoo! and Google where nearest groups of matches is identified to groups of keywords indicated by the user. For such type of applications, documents are indicated in the pattern of inverted index. Also, other access techniques like signature file exist, however the inverted image appears to have develop into the technique of selection in the IR domain. For every word 'w', it has all document identifiers, such that the related documents have it. Additionally, Meta information about word frequency, document length or position can be stored with every identifier. For every user query, it is adequate to analyze the Document (ID's) in inverted lists similar to the words in target or query. Similarity search has shown to be an attractive issue in a text domain due to the oddly huge dimensionality of the issue than that of documents size.

Signature extraction methods

They are superimposed coding technique, Compression technique and Word signature method.

Superimposed coding technique

This technique is akin to the compression technique. But every word is mapped or correlated into an individual mark. Keyword is considered as the fundamental element and record is considered as a group of individual elements or keyword, in any database. Signature of Record, Query and Keyword are encoded as binary image that characterize the core of them. Record is considered as a document as well as individual element is considered as a keyword, in Library database. With the help of superimposed coding method, the record signature is quantified through superimposing of its individual elements' signature. If keyword contains above one word, then the keyword signature is superimposing of words' signature.

Compression technique

In this compression technique, the document is separated into logical blocks i.e. a part of text that have a stable quantity of words). The design is to employ a (huge) bit vector of size zero 'n' for every block and then hash that particular block into a single bit positions that are set as 1. The latent with vector sparse may be compressed. After that, the resulted bit vectors are considered to be logically 'O' Red in order to obtain the document signatures. In the current contest, the super imposed coding method for measuring the signature is employed because it provides high-quality retrieval performance with effective storage. In this technique, all signatures are in same size.

Word signature method

In this technique, a single word of document is mixed up into the bit word signatures of length 'n'. These word signatures or patterns are concatenated to form the document signature.

Conclusion

Similarity search is one of the effective information retrieval techniques that have turn into a basic computational task in various application regions, including Digital Libraries, Data Mining, Data Compression, Statistical data analysis, Multimedia information retrieval and Bio-medical databases. In such kind of atmospheres, an accurate match has small meaning, and distance or proximity (Similarity or Dissimilarity) concepts are usually more successful for searching. These digital libraries have unstructured metadata, which consists of title, authors, keywords, topics, categories etc., These metadata information are error prone due to typographical, machine or manual, and hence to retrieve the correct information from the database is difficult.

Compliance with ethical standards

Conflict of interest This paper has not communicated anywhere till this moment, now only it is communicated to your esteemed journal for the publication with the knowledge of all co-authors.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Pranav, A., and Chauhan, S., Efficient focused web crawling approach for search engine. *International Journal of Computer Science and Mobile Computing* 4(5):545–551, 2015.
2. Lu, J. and Callan, J., Federated Search of Text-Based Digital Libraries in Hierarchical Peer-to-Peer Networks, European Conference on Information Retrieval, pp. 1–15, 2005.
3. Kovacevic, A., Devedzic, V., and Pocajt, V., Using data mining to improve digital library services. *The Electronic Library* 28(6):829–843, 2010.
4. Knoll, A. et al., An Integrated Approach to Semantic Evaluation and Content-Based Retrieval of Multimedia Documents, 1998, Retrieved on 18th April 2018 and from https://pdfs.semanticscholar.org/05d1/37266863f665135b060f47719f9ced5f7e1c.pdf.
5. Waugh, L., Tarver, H., Phillips, M. and Alemneh, D., Comparison of Full-text Versus Metadata Searching in an Institutional Repository: Case Study of the UNT Scholarly Works, 2015 Retrieved on 18th April 2018 and from https://arxiv.org/ftp/arxiv/papers/1512/1512.07193.pdf.
6. Memar, S., Affendey, L. S., Mustapha, N., Doraisamy, S. C., and Ektefa, M., An integrated semantic-based approach in concept based video retrieval. *Multimedia Tools and Applications* 64(1):77–95, 2013.
7. Wulandari, L. et al., User requirements analysis for digital library application using quality function deployment. *J. Phys.: Conf. Ser.* 818:1–11, 2017.
8. Calhoun, K. (2013) Key themes and challenges in digital libraries, Faecet Publishing, pp.1–33.