Asian Journal of Information Management

ISSN 1819-334X

Academic Journals Inc.

www.academicjournals.com

Asian Journal of Information Management 4 (1): 12-18, 2010
ISSN 1819-334X

Contextual Information Retrieval for Multi-Media Databases with Learning by Feedback Using Vector Space Model

V. Prasannakumari
Systems Analyst, Tek-Tools Software Solutions Pvt. Ltd., Chennai, India

Abstract: Information retrieval from huge databases has been an area of research to improve efficiency in terms of performance and precision of results returned. This study deals with contextual information retrieval from multimedia databases which have feature descriptors and metadata for the data items in it. This approach uses vector space method for IR and uses a stemming process to throttle feature descriptors to root words which in turn increases efficiency. Learning by feedback enables the database to accommodate more feature descriptors related to the data and builds it as a better described database. We have confined our study to IR alone excluding the considerations of data representation and storage techniques. Proposed method provides efficient search results and adds relevant contextual information to data for improvisation.

Key words: IR, MMDB, information retrieval, multimedia database, vector space, stemming, learning by feedback

INTRODUCTION

Storage and retrieval of multimedia data through computers have grown tremendously in the recent past. As databases provide consistency, concurrency, integrity, security and availability of data, processing different multimedia-related applications in databases proves advantageous. From an user perspective, the databases should provide functionalities for the easy manipulation, query and retrieval of highly relevant information from huge collections of stored data. Information Retrieval System (IRS) is a system used to store items of information that need to be processed, searched and retrieved corresponding to a user's query.

MULTI-MEDIA DATA-BASE-MMDB

Multimedia data means data as digital images, audio, video, animation and graphics together with text. A Multi Media Data Base is the one that can store and process several different types of information pertaining to the actual media data. They are:

- **Media Data:** Is the actual data representing images, audio, video that are captured, digitized, processed, compressed and stored
- **Media Format Data:** Is the information pertaining to the format of the media data after it goes through the acquisition and processing phases. This consists of information such as resolution, frame rate, etc.

- **Media Keyword Data:** Contains the keyword descriptions, usually related to the origin of the media data. For example, for a photograph image, this might include the date, time and place of capture, the person who took, etc. This is also called as content descriptive data
- **Media Feature Data:** Contains the features derived from the media data that characterizes the media content. For example, this could contain information about the features present in the media data. This is also referred to as content dependent data

The last three types that describe different aspects of media data are called meta data. The media keyword data and media feature data are used as indices for searching purpose. The media format data is used to present the retrieved information.

INFORMATION RETRIEVAL SYSTEM-IRS

An IRS is basically constituted by three main components, whose composition is as follows:

Representation in MMDB

This component stores the documents and the representations of their information contents. An indexer module, which automatically computes a representation for each document by extracting the document features is associated to this component. Meta Data serves as the content identifier when indexed.

Query for MMDB

This allows the users to express their information needs and presents the relevant data retrieved by the system. A query language that collects the rules to generate legitimate queries and procedures to select the relevant documents is the core for this component.

A typical query to fetch all images having a white car should be like,

```
SELECT * FROM PhotoTable WHERE Contains(car, white);
```

Comparing the Fetched Data

This component should evaluate the degree to the document, which satisfy the requirements expressed in the query. Data relevant and nearest to the requirements are presented to the user.

Learning by Feedback

This component enables the IR to update itself with more specific and appropriate index terms for the data from the interactions made by the user. As the user revises and provides more feature details for the data he is looking for, the resultant data is updated with the index terms it is missing.

LITERATURE SURVEY

De Vries and Henk Blanken (1998) provides a detailed study on the relationship between IR and multimedia databases.

Maybury's (1997) tutorial provides an overview of intelligent information access technologies: information retrieval, summarization, information extraction, text clustering and question answering.

Ferbers (1996) work dealt about accessing multimedia documents by knowledge discovery methods and intelligent retrieval and he presented an architecture called MAGIC-Multimedia-based Automatic-Generation of Indexes and Clusters.

Wong *et al.* (2005) proposed a novel query routing strategy called GAroute based on the query propagation model. By giving the current P2P network topology and relevance level of each peer, Garoute returns a list of query routing paths that cover as many relevant peers as possible. He model this as the Longest Path Problem in a directed graph which is NP-complete and we obtain high quality (0.95 in 100 peers) approximate solutions in polynomial time by using Genetic Algorithm (GA).

Henrik Bulskov (2006) explored Ontology-based Information Retrieval by mapping OntoLog Expressions into the Ontology. His method uses simple fuzzy retrieval for query evaluation and weighted shared nodes for comparison.

Wen *et al.* (2004) analyzed Probabilistic Model for Contextual Retrieval for applications such as mobile search, personalized search, PC troubleshooting. Their study proved that query log is the key to build effective contextual retrieval models.

Jan and Kostial (2003) presented an ontology-based approach to information retrieval. It was based on a domain knowledge representation schema in form of ontology resources were retrieved based on the associations and not only based on partial or exact term matching.

Redon *et al.* (2007) worked on getting a context based search platform that is a self-learning software system that enables the user to search for the Knowledge Elements in engineering.

IR MODELS

- Exact Match models
  - String matching
  - Boolean
- Best (partial match) models
  - Vector space
  - Probabilistic
  - Logic (Plausible inference)
  - Language modeling

Vector Space Model

In this model, a document is viewed as a vector in n-dimensional document space (where n is the number of unique terms used to describe contents of the documents in the collection) and each term represents one dimension in the document space. A query is also treated in the same way and constructed from the terms and weights provided in the query. Document retrieval is based on the measure of similarity between the query and the documents. This means that documents with a higher similarity to the query are presented to be more relevant to it and should be retrieved by the IRS in a top position in the list of retrieved documents. In this method, the retrieved documents can be presented in an order to the user with respect to their relevance to his information needs.

- Goal of IR is to present the user with data that is most similar to query, in order of similarity

- Similarity is defined as closeness in the concept (vector) space
- Uncertainty in IR is in the degree of match between concept space and query, arises from uncertainty in representation of each
- Advantages of Vector Model
  - Straightforward computation of closeness/relevance
  - Simple query formulation (bag of words)
  - Intuitively appealing
  - Effective

PROPOSED MODEL

Multimedia data is stored with all the feature descriptors available from its metadata. All feature descriptors are subjected to a stemming process defined.

Stemming Algorithm

Stemming is a process of removing and normalizing common suffixes, plural forms to derive a common root for words. More stringent stemming process brings down the group of related words into a common root. This is implemented by defining all the possible descriptors and their suggested root word. For example,

Stemming (Transport, shipping, Transfer, Commute, Carry, Move, Send) = Transport

All these related words are stemmed to a root word as "Transport". Extensively populated the list of possible words and their root words in stemming algorithm makes it more efficient. Stemming can be populated from Thesauruses for information retrieval that are typically constructed by information specialists and have their own unique vocabulary defining different kinds of terms and relationships.

Querying

User is encouraged to open a new dialogue for every new requirement and is expected to provide as many feature descriptors for the data they expect to retrieve. The descriptors in the user query are normalized by the same stemming process algorithms.

Ranking and Retrieval

A function that computes relevance then compares each data in the MMDB for the features list provided by the user and computes a rank for each of them using the below algorithm.

- **Step 1:** Normalize feature descriptors in user query (QD) using the stemming process.
- **Step 2:** For each data item k and its feature descriptors FD(k) in the database, computer rank(k) as:

if (QD exists in FD(k) then rank(k)++

Having computed ranks of relevance for all the data items in the database, the item having highest rank is presented to the user. If many have the highest rank, top N items can be presented to the user for choosing what they looked for.

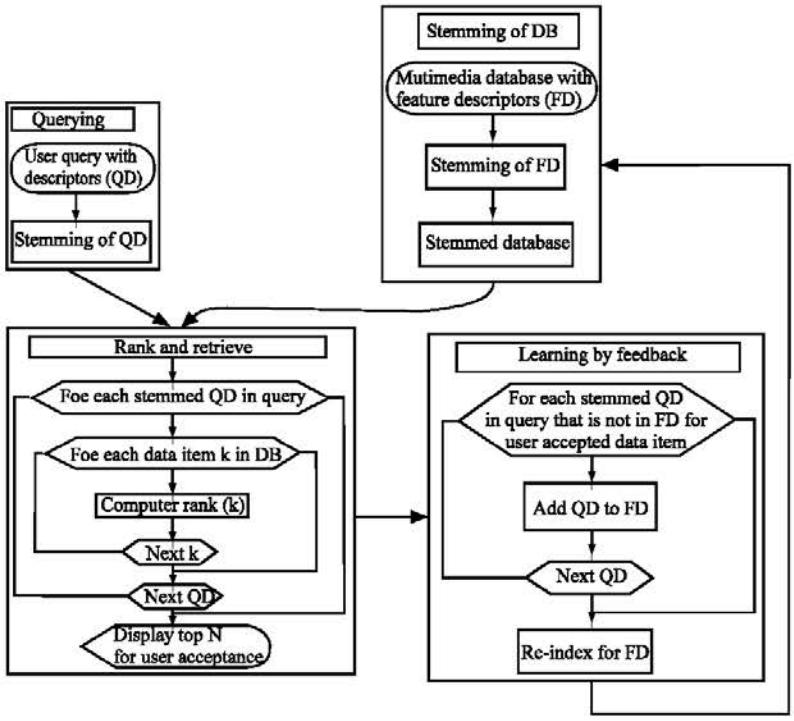


Fig. 1: Framework of the model

Learning by Feedback

If the user accepts any of the result presented which doesn't have any of the feature descriptors that was mentioned in the user query, still being retrieved by the relevance it had for other descriptors, needs to be added to the feature description list for data item retrieved. Framework of the proposed model is shown in Fig. 1.

On any update on descriptors, the database is automatically re-indexed. This enables the database to respond better for similar queries in future.

SIMULATED RESULTS

To simulate the method, let us assume the below and trace how the method would respond.

Stemmed Database

Sample set of images in the PhotoTable as shown below:

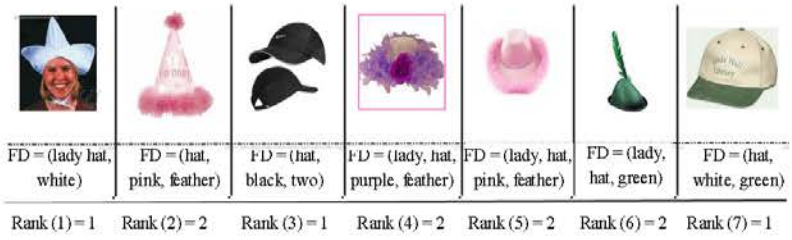
FD = (lady hat, white	FD = (hat, pink, feather)	FD = (hat, black, two)	FD = (lady, hat, purple, feather)	FD = (lady, hat, pink, feather)	FD = (lady, hat, green)	FD = (hat, white, green)

User Query

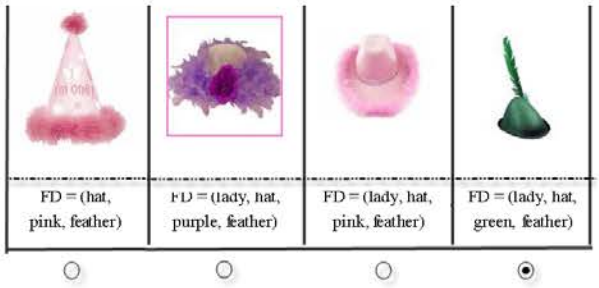
User wants to get the image of "alpine, feather, cap"
Select * from PhotoTable where contains(alpine, feather, cap)
QDs→(alpine, feather, cap)
Stemming of the QDs→(alpine, feather, hat)→
Note: (cap, hat, headdress) are stemmed into (hat)

Rank and Retrieve

In this module every image is processed for existence of contextual feature descriptor and their respective ranks are computed. Thus the ranks would be as below:



Having computed the ranks of relevance, results having high relevance ranks are presented to the user for acceptance as given below:



Learning by Feedback

User selected data is now analyzed for comparing FD and QD. QD-alpine missing in FD is detected and added to FD of the data accepted by user as shown below:

If the same query is submitted next time, rank for this data item would be 3, providing highest rank for its retrieval.

Above illustration explains that the user requirement is retrieved efficiently from the multimedia database and provides top N chances for him to choose the best out of them.

Feature description addition through feedback enables to build stronger FD list for the accepted item thus enhancing it for future.

This approach is more efficient and enhanced from all the previous studies in the way that it combines ranking and learning by feedback approach. Unlike many of the previous works, ours is a loss-less watermark for the data and its metadata.

CONCLUSION

This method is a simplest alternative yet efficient method for contextual retrieval of multimedia data from a MMDB for any user requirement. As this subjects the feature descriptors to the strong stemming process, the list of descriptors can be kept to minimum, thus maximizing the usability and performance. Learning by feedback educates the DB for better relevance ranking which will retrieve better and nearest matching data for the same query next time.

REFERENCES

De-Vries, A.P. and M. Henk Blanken, 1998. The Relationship between IR and Multimedia Databases. Autrans, France.

Ferber, R., 1996. Digital libraries research at GMD-IPSI: Accessing multimedia documents by knowledge discovery methods and intelligent retrieval. http://www.ercim.org/publication/ws-proceedings/DELOS1/ferber.pdf.

Henrik Bulskov, 2006. Ontology-based information retrieval. http://coitweb.uncc.edu/~ras/Ontology-IR.PPT.

Jan, J.P. and I. Kostial, 2003. Ontology-based information retrieval. Proceedings of 14th International Conference on Information and Intelligent systems (IIS), (ICIIS. 03), Varazdin, Croatia, pp: 23-28.

Maybury, M., 1997. Intelligent Multimedia Information Retrieval. AAAI Press/MIT Press, London.

Redon, R., A. Larsson, R. Leblond and B. Longueville, 2007. VIVACE context based search platform. Proceedings of 6th International and Interdisciplinary Conference, Aug. 20-24, Roskilde, Denmark, pp: 397-410.

Wen, J.R., N. Lao and W.Y. Ma, 2004. Probabilistic model for contextual retrieval. Proceedings of the 27th Annual International ACM Sigir Conference on Research and Development in Information Retrieval, (AISCRDIR. 04), Sheffield, UK, pp: 57-63.

Wong, W.Y., T.P. Lau and I. King, 2005. Information retrieval in P2P networks using genetic algorithm. Proceedings of International World Wide Web Conference, (IWWWC. 05), Chiba, Japan, pp: 922-923.