ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

A Study on Creating Assessment Model for Miniature Question Answer Using Nearest Neighbor Search Keywords

L.Mary Immaculate Sheela[1], R.J.Poovaraghan[2]

M.Tech Student, Dept. of CSE, SRM University, Chennai, India[1]

Assistant Professor (OG), Dept. of CSE, SRM University, Chennai, India[2]

ABSTRACT: We introduce a new Assessment model for nearest neighbour search keywords algorithm. The algorithm builds a nearest neighbour model. There are limited types of knowledge that can be assessed by multiple choice tests. In multiple choice questions a student can have simply select a random answer and still have a chance of receiving a mark for it. In a short answer question, the student types in a word or phrase or sentences in response to a question. The answer could be a word or a phrase, but it must match one of our acceptable keywords exactly. It's a good idea to keep the required answer as short as possible to avoid missing a correct answer that's phrased differently. This feature allows a user to create, preview, and edit questions in a database of question categories. In this paper we are going to compare the entire template model produced by both training and testing phase. Measure the similarities between the two models and identify the quality and efficiency of the result using nearest neighbour keyword. The experimental results show what aspects are important for short answer question type classification in terms of both effectiveness and efficiency. We believe that the assessment model findings from this study will be useful in real-world student problems.

Keywords: Nearest Neighbour(NN), Multiple Choice Question,, Parsing, Feature Matrix, Template model.

I. INTRODUCTION

The simplest solution to the NNS problem is to compute the distance from the query point to every other point in the database, keeping track of the "best so far". Current research on queries based on the followings: nearest neighbor queries, range queries [1], and spatial joins [7], [6].We will discuss some of the existing method available for nearest-neighbor (NN) search. Nearest neighbor (NN) retrieval, involve solely conditions on objects' geometric properties. Today several trendy applications demand novel styles of queries that aim to seek out objects satisfying each an abstraction predicate, and a predicate on their associated texts.

A. Linear search

This algorithm, sometimes referred to as the naive approach, has a running time of $O(Nd)$ where N is the cardinality of S and d is the dimensionality of M. There are no search data structures to maintain, so linear search has no space complexity beyond the storage of the database. Naive search can, on average, outperform space partitioning approaches on higher dimensional spaces.

B. k-nearest neighbour

K-nearest neighbor search identifies the top k nearest neighbors to the query. This technique is commonly used in predictive analytics to estimate or classify a point based on the consensus of its neighbors. K-nearest neighbor graphs are graphs in which every point is connected to its k nearest neighbors.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

C. Approximate nearest neighbor

In some applications it may be acceptable to retrieve a "good guess" of the nearest neighbor. In those cases, we can use an algorithm which doesn't guarantee to return the actual nearest neighbor in every case, in return for improved speed or memory savings. Often such an algorithm will find the nearest neighbor in a majority of cases, but this depends strongly on the dataset being queried. Algorithms that support the approximate nearest neighbor search include locality-sensitive hashing, best bin first and balanced box-decomposition tree based search.

II. RELATED WORK

The Keyword search has been well studied for years due to its importance to commercial search engines. Various types of keyword queries have been proposed. These related works can be categorized from two phases first, we introduce the works with Training which requires the input documents' associated with or contained in a query region. The query keywords contain a priority. The result documents are ranked based on certain criteria. The query processing contains two stages, Parsing and Frequency finder using Data Mining technique. The textual attribute is represented by a list of keywords stored in a Feature Matrix. After the textual parsing, second phase testing will take place to build an inverted list for each keywords and the query are inserted in the template model. To improve performance, S2I distinguishes between frequent and infrequent keywords. The model contains following modules.

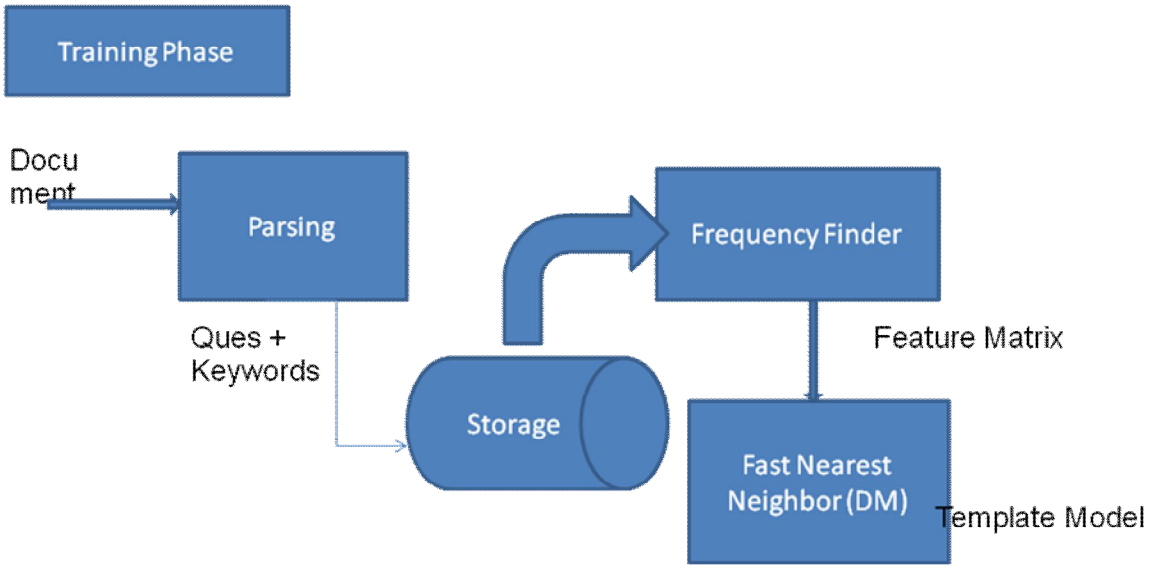


Fig:1 FNN Block Diagram phase1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

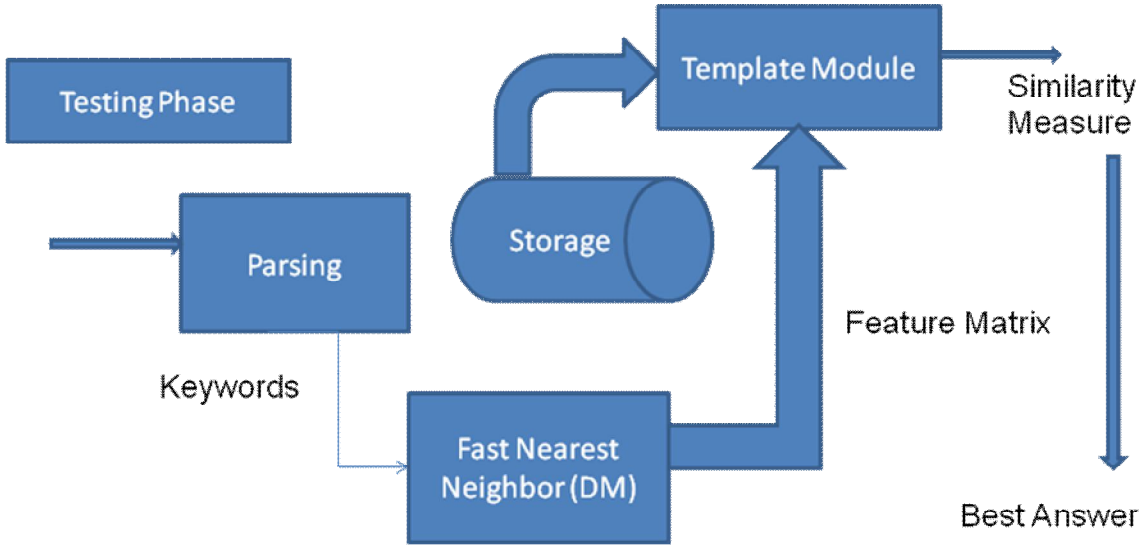


Fig:2 FNN Block Diagram phase2

A. *Feature Matrix*

The table is essentially a map whose key is a keyword and the value contains a pair of fields. Our data file contains a sequence of fixed-size pages. Each page is parsed into a fixed number of slots, one slot for one query We discard the requirements of ranking order and page contiguity in inverted index so that different keyword cells in the same list can be stored in different pages and updated concurrently.

B. *Query Processing*

We first present the overview of the query processing algorithm. Then, we explain different parsing techniques based on Fast Nearest Neighbor. Since all the keywords follow the same mechanism, our query processing algorithm starts from the root adopts a top-down search strategy to access child cells. Each cell is a keyword.

C. *Experiment Evaluation*

In this section, we compare the performance of Training and testing phase of Fast Nearest Neighbor Block diagram [fig1]. We do not show experiment results of frequency finder other than to show little improvement in query processing performance to build the template model. All of the indexes were implemented in Java we sorted these queries based on the frequency of keywords.

D. *Training Phase*

1. Input - type question and answer with keywords and store it in the direct access storage.
2. Parsing - create Inverted index and stored separately in a file. It is used to separate the keywords from the answer typed by the user.
3. FNN - Identify the keyword and create a template model for each question answer using parser.

E. *Testing Phase*

1. Query – In this module the user has to select the question and type the answer in phrase or word

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

in response to a question.

2. Parsing – this method used to separate the keywords from the answers typed by the user.
3. FF (Frequency Finder) – This module used to identify the number of keywords in the answer are repeated and make a count for each keyword.
4. Template module - Create a template module for answer typed by the user and compare the keyword Matching.
5. SMM - Similarities Measurement Module is used to find the similarities between the training and testing template module keywords.

III. LITERATURE WORK PRILIMINARIES

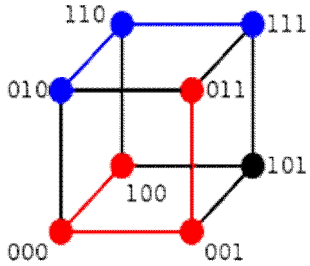
There is much evidence to show that assessment is the driving force behind student learning (Beevers et al (1991), Brown et al (1997)).[14] [15]. It is increasingly the case that students are becoming ever more strategic in their study habits and they are reluctant to undertake work which does not count towards their final grade. The method of assessment is classified into

• Diagnostic assessment --- tests which aim to determine a student's prior knowledge of a subject area or preparedness for a course of study;
• Self-assessment --- tests which show students whether they understand particular concepts and terminology where feedback is given but results are not recorded;
• Formative assessment --- an assessment for which the primary purpose is to promote learning by providing feedback, but which does not count towards the final mark or grade though marks may be recorded for information only or to bring the teacher into the process; and
• Summative assessment --- assessments which count towards the final mark or grade.

Currently the best solution to such queries is based on the IR2-tree, which, as shown in this project, has a few deficiencies that seriously impact its efficiency. Motivated by this, we develop a new access method called the spatial inverted index that extends the conventional inverted index to cope with multidimensional data, and comes with algorithms that can answer nearest neighbor queries with keywords in real time.

A. Hamming Distance

In information theory, the Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different.



Red path has a hamming distance of 3(100 requires all 3 bits changed to get to 011)
Blue path has a hamming distance of 2(010 requires all 2 bits changed to get to 111).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

B. IR2 Tree

In the past years, the keyword search is used in relational databases. It is until recently that attention was diverted to multidimensional data [3], [4]. The best method for nearest neighbor search with keywords is R-tree [2], a popular spatial index, and Signature file [2], an effective method for keyword-based document retrieval. By doing so they develop a structure called the IR2 -tree [3], which has the strengths of both R-trees and Signature files. Like R-trees, the IR2 -tree preserves objects spatial proximity, which is the key to solving spatial queries efficiently. On the other hand, like signature files the IR2 -tree is able to filter a considerable portion of the objects that do not contain all the query keywords, thus significantly reducing the number of objects to be examined.
The IR2 -tree is an R-tree where each (leaf or non leaf) entry E is augmented with a signature that summarizes the union of the texts of the objects in the sub tree of E. On conventional R-trees, the best - first algorithm is a well-known solution to NN search.

C. Drawbacks of the IR2 –Tree

The IR2 -tree is the first access method for answering NN queries with keywords. As with many pioneering solutions, the IR2 -tree also has a few drawbacks that affect its efficiency. The most serious one of all is that the number of false hits can be really large when the object of the final result is far away from the query point, or the result is simply empty. A serious drawback of the R-tree approach is its space cost. Notice that a point needs to be duplicated once for every word in its text description, resulting in very expensive space consumption.

IV. EXPERIMENTS AND RESULTS

Table1 Subject : Database Management

QNO	Question	Answer	Keywords
1	Define database management system?	Database management system (DBMS) is a collection of interrelated data and a set of programs to access those data.	Interrelated, Data, set, programs.
2	Define instance and schema?	Instance: Collection of data stored in the data base at a particular moment is called an Instance of the database. Schema: The overall design of the data base is called the data base schema.	collection , data, database, overall, design, Structure.

Sample Coding for Parsing:

The user enters text into a search box. This class is used to parse that text into specific keywords (or tokens). It eliminates common words, and allows for the quoting of text, using double quotes. Strings in Java can be parsed using the split method of the String class.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

Input: Define Database Management System.

String answer = " Collection of Interrelated Data or set of programs **";**
String delims = "[ ]+";
String[] tokens = answer. Split(delims);

Output (keywords): collection, of ,Interrelated, data ,or, set , of, programs

V. CONCLUSION

This paper is the first work on short question type classification. We also study the usefulness of several components of questions in different subject representations. The main conclusion that I want to represent from the discussion in this paper is that the accuracy and efficiency of keyword measure, which we can try to solve in different ways but which always requires a text parsing methods. Inverted files are used to index text. The indices are appropriate when the text collection is large and semi-static. If the text collection is volatile online searching is the only option. Some techniques combine online and indexed searching. Most of the data structures and techniques proposed since the initial inception of the NNS problem have not been extensively compared with each other, making it hard to gauge their relative performance. Our approaches obtain the best results in terms of the trade-off between accuracy and keyword. We have reported some new findings from our experimental study, and the results reported in this paper could have attracted the attention of student community.

REFERENCES

1. Clarkson, Kenneth L, 'Fast algorithms for the all nearest neighbors problem', 24th IEEE Symp. Foundations of Computer Science, (FOCS '83), pp. 226–232, doi:10.1109/SFCS.1983.16.
2. C. Faloutsos and S. Christodoulakis.' Signature files: An access method for documents and its analytical performance evaluation', *ACM Transactions on Information Systems (TOIS)*, 2(4):267–288, 1984.
3. I. D. Felipe, V. Hristidis, and N. Rishe,'Keyword search on spatial databases', In *Proc. of International Conference on Data Engineering (ICDE)*, pages 656–665, 2008.
4.] .R. Hariharan, B. Hore, C. Li, and S. Mehrotra,'Processing spatialkeywor (SK) queries in geographic information retrieval GIR systems' In *Proc. of Scientific and Statistical Database Management (SSDBM)*, 2007.
5. X. Cao, G. Cong, and C. S. Jensen,' Retrieving top-k prestige-based relevant spatial web objects', *PVLDB*, 3(1):373–384, 2010.
6. Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma,' Hybrid index structures for location-based web search', In *Proc. of Conference on Information and Knowledge Management (CIKM)*, pages 155–162, 2005.
7. Jegou H, Douze M, Schmid C,' Product quantization for nearest neighbor search IEEE Trans',. Patt. Anal. Mach. Int. in press. [PubMed]
8. E. Chu, A. Baid, X. Chai, A. Doan, and J. Naught on. ,'Combining keyword search and forms for ad hoc querying of databases',In *Proc. of ACM Management of Data (SIGMOD)*, 2009.
9. Beyer K, Goldstein J, Ramakrishnan R, Shaft U.,'When is "nearest neighbor meaningful?', Proceedings of Database Theory—ICDT'99; Jerusalem, Israel. January 1999; pp. 217–235.
10. Clarkson, Kenneth L,'Fast algorithms for the all nearest neighbors problem', *24th IEEE Symp. Foundations of Computer Science (FOCS '83)*, pp. 226–232, doi:10.1109/SFCS.1983.16.
11.] Cayton, Lawerence , 'Fast nearest neighbor retrieval for bregman divergences', *Proceedings of the 25th international conference on Machine learning*: 112–119. 2008.
12. Bendersky, M., & Croft, W. B,' Discovering key concepts in verbose queries', In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval* (pp. 491–498). New York, NY, USA: ACM. 2008.
13. John Heywood, *Assessment in Higher Education*; Jessica Kingsley Publishers,London. Pages 350-372, include guidance about various psychometric properties of MCQs and how to calculate and deal with them. 2000.
14. Brown, G., with Bull, J., and Pendlebury, M, 'Assessing Student Learning in Higher Education', London: Routledge. 1997.
15. Beevers, C., Cherry, B., Foster M. and McGuire, G,'Software Tools for Computer Aided Learning in Mathematics', Ashgate Publishing Company.1991.

BIOGRAPHY

Mary Immacullate Sheela L is a M.Tech Student in Computer Science and Engineering Department, SRM University, Chennai, India. She received Master degree from St.Joseph's College, Trichy, India. Her research interests are Sensor Fusion Algorithms, Wireless Sensor Networks, Image Processing, Algorithms,etc.