Boolean Revival for Deft Protracted

[1]M.Srivani, [2]K.Vasanth Kumar, [3]P.Suresh Babu
[1]Dept. of CSE, Pydah College of Engineering & Tech., Visakhapatnam, AP, India
[2]Dept. of CSE & IT, Kaushik College , Visakhapatnam, AP, India
[3]Dept. of CSE, Kaushik College of Engineering, Visakhapatnam, AP, India

Abstract
This paper gives a modern research model for extracting patterns and relations visually from multidimensional binary data using monotone Boolean functions. With the growth of massive digital data archives, which are not necessarily organized in any order, the twin and complementary processes of information retrieval and data mining have emerged together as a particular important discipline within the information sciences. The object of information retrieval is to automatically search a data archive in order to respond to a user's query. The object of data mining, on the other hand, is to automatically process a data archive in order to find patterns that represent knowledge or, equivalently, information interesting to the user (not necessarily in response to a targeted query). Information retrieval and data mining invoke multidisciplinary techniques, including those from artificial intelligence, statistics, machine learning, pattern analysis, and others.

Keywords
Visual Data Mining, explicit data structure, Boolean data, Monotone Boolean Function, Hansel Chains, Binary Hypercube

I. Introduction

In recent years, there is rapid development of digital data made available for knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential step in the process of knowledge discovery in databases. Search service providers have an interest in delivering competitive effectiveness levels within the smallest possible resource cost. This is no less true in specialized search services dedicated to medical and legal literature, which are called upon to support complex queries by professional searchers, possibly with significant commercial or societal outcomes resting on the results of the search. In particular, although the number of queries submitted per day to biomedical search engines is orders of magnitude less than the number submitted to web-scale search systems (millions per day for PUBMED,1 rather than billions per day for free web search), such services are typically funded as public services rather than by advertising; the queries are often much more complex, involving dozens or even hundreds of terms; there is a great deal of reformulation and reevaluation; and the user evaluation process typically involves hundreds or thousands of answer documents rather than a mere handful. Ranked retrieval has been successfully deployed in a wide range of applications. The main advantages of ranking are the simplicity of querying, and that results are ordered by estimated relevance, so that query quality can quickly be assessed once the top few results have been inspected. Having the answers returned as a ranked list also gives users the ability to consciously choose the amount of effort they are willing (or able) to invest in inspecting result documents. However, Boolean retrieval has not been superseded, and is still the preferred method in domains such as legal and medical search. Advantages of Boolean retrieval include: . Complex information need descriptions: Boolean queries can be used to express complex concepts;. Compos ability and Reuse: Boolean filters and concepts can be recombined into larger query tree structures;. Reproducibility: Scoring of a document only depends on the document itself, not statistics of the whole collection, and can be reproduced with knowledge of the query; . Scrutability: Properties of retrieved documents can be understood simply by inspection of the query; and . Strictness: Strict inclusion and exclusion criteria are inherently supported, for instance, based on metadata. For these reasons, Boolean retrieval—and the extended Boolean variant of it that we pursue in this paperremains a critically important retrieval mechanism. For carefully formulated information needs, particularly when there are exclusion criteria as well as inclusion criteria, ranking over bags of words is not appropriate. As one particular example, recent results suggest that ranked keyword queries are not able to outperform complex Boolean queries in the medical domain [1].

Boolean queries have the disadvantage of being harder to formulate than ranked queries, and, regardless of the level of expertise of the user, have the drawback of generating answer lists of unpredictable length. In particular, changes in the query that appear to be small might result in disproportionately large changes in the size of the result set. This is a problem that even expert searchers struggle with, adding and removing terms and operators until a reasonably sized answer set is retrieved, potentially even at the expense of retrieval effectiveness. Only when the answer set is of a manageable size can the searcher begin to invest time in examining its contents. Extended Boolean Retrieval (EBR) models, such as the pnorm model, seek to rank on the basis of Boolean query specifications [2-3]. They generate a list of top-k answers that can be extended if required, without necessarily sacrificing detailed control over inclusion and exclusion of terms and concepts. But EBR queries are slow to evaluate, because of their complex scoring functions [4]; and none of the computational optimizations available for ranked keyword retrieval have been applied to EBR. In particular, approaches that involve nonexact methods, such as quantized impact-ordered indexes or index pruning [5], do not satisfy all of the requirements listed above.

II. Current Research Work

In our Existing System, A significant amount of work has been devoted to the evaluation of top-k queries in databases. Provide a survey of the research on top-k queries on relational databases. This line of work typically handles the aggregation of attribute values of objects in the case where the attribute values lie in different sources or in a single source. For example, Bruno etc. Consider the problem of ordering a set of restaurants by distance and price. They present an optimal sequence of random or sequential accesses on the sources (e.g., Zagat for price and Mapquest for distance) in order to compute the top- k restaurants. Since the concept of

top-k is typically a heuristic itself for locating the most interesting items in the database, Theobald et al. Describe a framework for generating an approximate top-k answer, with some probabilistic guarantees.

In our work, we use the same idea; the main and crucial difference is that we only have "random access" to the underlying database (i.e., through querying), and no "sorted access." Theobald et al. assumed that at least one source provides "sorted access" to the underlying content.

III. Implementation of Research Work

We present a scoring method for EBR models that decouples document scoring from the inverted list evaluation strategy, allowing free optimization of the latter. The method incurs partial sorting overhead, but, at the same time, reduces the number of query nodes that have to be considered in order to score a document. We show experimentally that overall the gains are greater than the costs. We adopt ideas from the max-score and wand algorithms and generalize them to be applicable in the context of models with hierarchical query specifications and monotonic score aggregation functions. Further, we show that the p-norm EBR model is an instance of such models and that performance gains can be attained that are similar to the ones available when evaluating ranked queries. Term-independent bounds are proposed, which complement the bounds obtained from max-score. Taken alone, term-independent bounds can be employed in the wand algorithm, also reducing the number of score evaluations. Further, in conjunction with the adaption of max-score, this novel heuristic is able to short-circuit the scoring of documents.

- Complex information need descriptions: Boolean queries can be used to express complex concepts.
- Composability & Re-use: Boolean filters and concepts can be recombined into larger query tree structures.
- Reproducibility: Scoring of a document only depends on the document itself, not statistics of the whole collection, and can be reproduced with knowledge of the query.
- Scrutability: Properties of retrieved documents can be understood simply by inspection of the query.
- Strictness: Strict inclusion and exclusion criteria are inherently supported, for instance, based on metadata.

A. Module Description

1. SP Mining.
2. PTM
3. IP Evolving
4. User Interface Design

1. SP Mining

In this module we generate a frequent sequential pattern is a maximal sequential pattern if there exists no frequent sequential pattern. The length of sequential pattern indicates the number of words (or terms) contained in pattern. A sequential pattern which contains n terms extracted from given set of documents. Here we take set of documents as input we generate nterm sequences.

2. PTM

In this module we present a pattern-based model PTM (Pattern Taxonomy Model) for the representation of text documents. Pattern taxonomy is a tree-like structure that illustrates the relationship between patterns extracted from a text collection. An example of pattern taxonomy. (i.e., maximum sequential patterns). Once the tree is constructed, we can easily find the relationship between patterns. The next step is to prune the meaningless patterns in the pattern taxonomy.

3. IPEvolving

In this module we take positive documents and negative documents and we adjust the term weights based on term weight of positive document and negative document. Using this technique we can increase maximum likelihood event one documents having more overlapping terms and less content of the document we get accurate results.

4. User Interface Design

In this module we design an user interface to operate with the system easily in order to brose the documents and to training, testing and prediction.

B. Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

VI. Design Analysis

A. Networking

1. TCP/IP Stack

The TCP/IP stack is shorter than the OSI one:

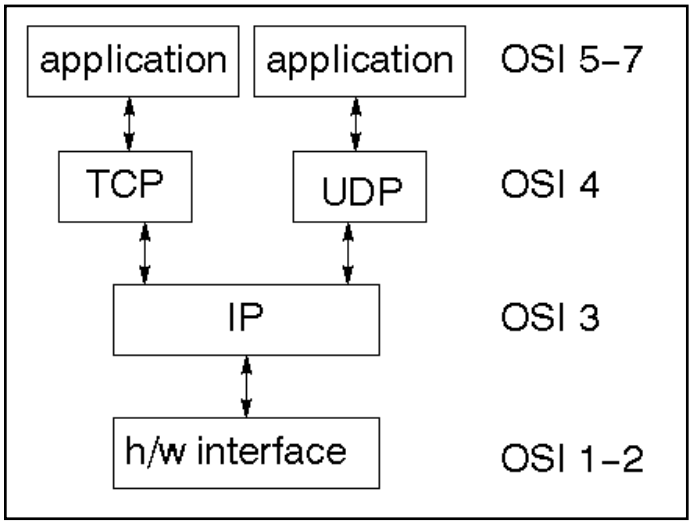

Fig. 1:

TCP is a connection-oriented protocol; UDP (User Datagram Protocol) is a connectionless protocol.

2. IP datagram's

The IP layer provides a connectionless and unreliable delivery system. It considers each datagram independently of the others. Any association between datagram must be supplied by the higher layers. The IP layer supplies a checksum that includes its own header. The header includes the source and destination addresses. The IP layer handles routing through an Internet. It is

also responsible for breaking up large datagram into smaller ones for transmission and reassembling them at the other end.

3. TCP
TCP supplies logic to give a reliable connection-oriented protocol above IP. It provides a virtual circuit that two processes can use to communicate.

4. Internet Addresses
In order to use a service, you must be able to find it. The Internet uses an address scheme for machines so that they can be located. The address is a 32 bit integer which gives the IP address. This encodes a network ID and more addressing. The network ID falls into various classes according to the size of the network address.

5. Network Address
Class A uses 8 bits for the network address with 24 bits left over for other addressing. Class B uses 16 bit network addressing. Class C uses 24 bit network addressing and class D uses all 32.

6. Subnet Address
Internally, the UNIX network is divided into sub networks. Building 11 is currently on one sub network and uses 10-bit addressing, allowing 1024 different hosts.

7. Host Address
8 bits are finally used for host addresses within our subnet. This places a limit of 256 machines that can be on the subnet.

8. Total Address

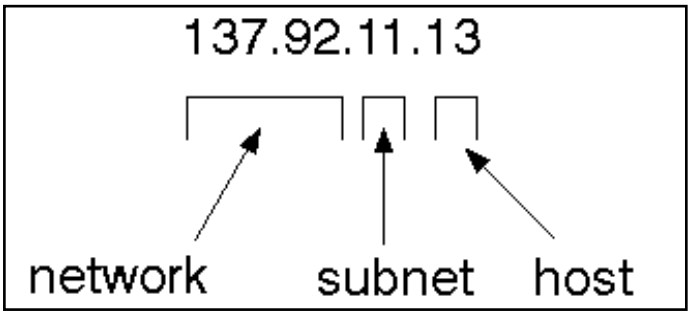

Fig. 2:

The 32 bit address is usually written as 4 integers separated by dots.

9. Port Addresses
A service exists on a host, and is identified by its port. This is a 16 bit number. To send a message to a server, you send it to the port for that service of the host that it is running on. This is not location transparency! Certain of these ports are "well known".

10. Sockets
A socket is a data structure maintained by the system to handle network connections. A socket is created using the call socket. It returns an integer that is like a file descriptor. In fact, under Windows, this handle can be used with Read File and Write File functions.

```
#include <sys/types.h>
#include <sys/socket.h>
int socket(int family, int type, int protocol);
```

Here "family" will be AF_INET for IP communications, protocol will be zero, and type will depend on whether TCP or UDP is used. Two processes wishing to communicate over a network create a socket each. These are similar to two ends of a pipe - but the actual pipe does not yet exist.

11. JFree Chart
JFreeChart is a free 100% Java chart library that makes it easy for developers to display professional quality charts in their applications. JFreeChart's extensive feature set includes: A consistent and well-documented API, supporting a wide range of chart types; A flexible design that is easy to extend, and targets both server-side and client-side applications; Support for many output types, including Swing components, image files (including PNG and JPEG), and vector graphics file formats (including PDF, EPS and SVG); JFreeChart is "open source" or, more specifically, free software. It is distributed under the terms of the GNU Lesser General Public Licence (LGPL), which permits use in proprietary applications.

V. System Desugn and Testing
The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

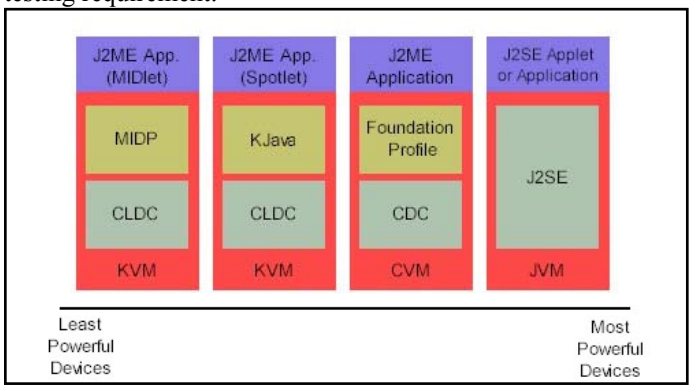

Fig. 3: General J2ME Architecture

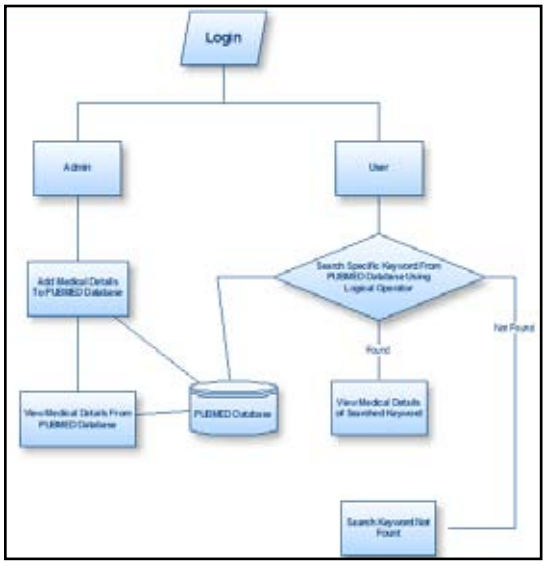

Fig. 4: Data Flow Diagram Of System Design

A. Implementation

1. Modules

- Using Boolean Condition (AND)
- Using Boolean Condition (OR)
- Using Boolean Condition (NOT)
- Top k-Query Search

B. Modules Description

1. Using AND Condition

We define the novel problem of applying ranking on top of sources with no ranking capabilities by exploiting their query interface. For instance, if the user query is Q= [anemia, diabetes, sclerosis], then we can submit to the data source queries q1 = [anemia AND diabetes AND sclerosis], q2 = [anemia AND diabetes AND NOT sclerosis], q3 = [diabetes OR sclerosis], and so on. The returned results are guaranteed to match the Boolean conditions but the documents are not expected to be ranked in any useful manner.

2. Using OR Condition

We describe sampling strategies for estimating the relevance of the documents retrieved by different keyword queries. We present a static sampling approach and a dynamic sampling approach that simultaneously executes the query, estimates the parameters required for efficient query execution, and compensates for the biases in the sampling process. For instance, if the user query is Q= [anemia, diabetes, sclerosis], then we can submit to the data source queries q1 = [anemia AND diabetes AND sclerosis], q2 = [anemia AND diabetes AND NOT sclerosis], q3 = [diabetes OR sclerosis], and so on. The returned results are guaranteed to match the Boolean conditions but the documents are not expected to be ranked in any useful manner.

3. Using NOT Condition

We present algorithms that, given a user confidence input, retrieve a minimal number of results from the source through submitting high selectivity (conjunctive) queries, so that the user's confidence requirement is satisfied. For instance, if the user query is Q= [anemia, diabetes, sclerosis], then we can submit to the data source queries q1 = [anemia AND diabetes AND sclerosis], q2 = [anemia AND diabetes AND NOT sclerosis], q3 = [diabetes OR sclerosis], and so on. The returned results are guaranteed to match the Boolean conditions but the documents are not expected to be ranked in any useful manner.

4. Top K-Query Search

Our overall goal is to figure out during our querying process, how many of the top-k relevant documents we have retrieved and how many are still unretrieved in the database. Unfortunately, we cannot be absolutely certain about these numbers unless we retrieve and score all documents: an expensive operation. Alternatively, we can build a probabilistic model of score distributions and examine, probabilistically, how many good documents are still not retrieved. We describe our approach here.

VI. Conclusion

Having noted that ranked keyword querying is not applicable in complex legal and medical domains because of their need for structured queries including negation, and for repeatable and scrutable outputs, we have presented novel techniques for efficient query evaluation of the pnorm (and similar) extended Boolean retrieval model, and applied them to document-at-a-time evaluation. We showed that optimization techniques developed for ranked keyword retrieval can be modified for EBR, and that they lead to considerable speedups. Further, we proposed term independent bounds as a means to further short-circuit score calculations, and demonstrated that they provide added benefit when complex scoring functions are used. A number of future directions require investigation. Although presented in the context of document-at-a-time evaluation, it may also be possible to apply variants of our methods to term-at-a-time evaluation. Second, to reduce the number of disk seeks for queries with many terms, it seems desirable to store additional inverted lists for term prefixes (see, for example, Bast and Weber [19]), instead of expanding queries to hundreds of terms; and this is also an area worth exploration.

We also need to determine whether or not term-dependent bounds can be chosen to consistently give rise to further gains. As another possibility, the proposed methods could further be combined and applied only to critical or complex parts of the query tree. Finally, there might be other ways to handle negations worthy of consideration. We also plan to evaluate the same implementation approaches in the context of the inference network and wand evaluation models. For example, it may be that for the data we are working with relatively simple choices of term weights—in particular, strictly document-based ones that retain the scrutability property that is so important—can also offer good retrieval effectiveness in these important medical and legal applications.

References

[1] K. Aas, L. Eikvil,"Text Categorisation: A Survey", Technical Report Raport NR 941, Norwegian Computing Center, 1999.

[2] R. Agrawal, R. Srikant,"Fast Algorithms for Mining Association Rules in Large Databases", Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

[3] H. Ahonen, O. Heinonen, M. Klemettinen, A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections", Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[4] R. Baeza-Yates, B. Ribeiro-Neto,"Modern Information Retrieval", Addison Wesley, 1999.

[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, C. Gentile, "Kernel Methods for Document Filtering", TREC, trec.nist.gov/pubs/trec11/papers/kermit.ps.gz, 2002.

[6] N. Cancedda, E. Gaussier, C. Goutte, J.-M. Renders, "Word-Sequence Kernels", J. Machine Learning Research, Vol. 3, pp. 1059-1082, 2003.

[7] M.F. Caropreso, S. Matwin, F. Sebastiani,"Statistical Phrases in Automated Text Categorization", Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.

[8] C. Cortes, V. Vapnik,"Support-Vector Networks", Machine Learning, Vol. 20, No. 3, pp. 273-297, 1995.

[9] S.T. Dumais,"Improving the Retrieval of Information from External Sources", Behavior Research Methods, Instruments, and Computers, Vol. 23, No. 2, pp. 229-236, 1991.

[10] J. Han, K.C.-C. Chang,"Data Mining for Web Intelligence", Computer, Vol. 35, No. 11, pp. 64-70, Nov. 2002

Mrs.M.Srivani is a student of PYDAH College of Engineering & Technology, Gambheeram, visakhapatnam. She is pursuing his M.Tech [C.S.E] from this college and she received her MCA from Vignan institute of information technology affiliated to JNTU University, Kakinada in the year 2011. Her area of Interest Data warehousing and Data mining.

Sri.K.Vasanth Kumar, excellent teacher received MS (Information Systems & Applications) from Bharathidasan University, Tiruchirapalli, Tamil Nadu and M.Tech (CSE) from JNTU Kakinada, Andhra Pradesh is working as Associate Professor Department of CSE & IT, Kaushik college of Engineering. He has 7 years of teaching experience in various engineering colleges and 4 years of industrial experience. To his credit 1 International publications and 4 work shops . His area of Interest includes Datawarehousing and Network Security.

Sri. P.Suresh Babu, B.Tech,M.E., CSI Associate Professor Department of Computer Science & Engineering Kaushik College of Engineering, Visakhapatnam, Andhra Pradesh. Teaching Experience: 14 Years Industrial Experience: 4 Years
Sri.P.Suresh Babu, well known Author and excellent teacher Received B.Tech(CSE) from Acharya Nagarjuna university, GUNTUR, Andhra Pradesh and M.E (CSE) from Sathyabama university, Chennai is working as Associate Professor Department of Computer Science and Engineering, Kaushik college of Engineering, He is an active member of CSI. .He has 14 years of teaching experience in various engineering colleges and 4 years of industrial experience. To his credit 7 International publications, 2 national publications, 2 International conferences and 4 work shops . His area of Interest includes

.