Retrieval of Information Document Using TF-IDF Algorithms and Vector Space Model Representation

Dr C. SUBA,
Principal cum Assist Professor, Bharathi Women's Arts & Science College, Thatchur,
Kallakurichi District, Tamil Nadu, India

Abstract: Retrieval of information systems such as web interfaces services is a great technology in the web search services. In this paper, we present different approaches of information retrieval using document vector representation. The goal of information retrieval (IR) is to provide users with those documents that will satisfy their information need. Retrieval models can attempt to describe the human Process, such as the information need for interaction. Retrieval of information process has been a prominent and ongoing research in the field of natural language processing. Information Retrieval (IR) is searching for document or information in documents. Document can be text or multimedia and may reside on the web. The vector space mode is one of the classical and widely applied retrieval models to evaluate relevance of web page and efficient search for best documents. The retrieval of information consists of computing the Modified Cosine Coefficient Similarity Measure (MCCSM). To improve the quality of the search result returned by the internet which makes users have to look through a huge amount of links for the real answers, we utilized the high quality links Google produces and the Information Retrieval technology to implement a Question Answering (QA) system. This system analyzes and downloads the text contents from the relevant web pages Google searches based on the users' questions to build a dynamic knowledge collection; retrieves the relevant passages from the collection and sends the ranked passages back. We utilized the high quality links Google produces and the Information Retrieval technology to implement a Question Answering (QA) system. In this paper, we present retrieval of document also involves the TF-IDF algorithm and Vector Space Model for the document indexing. We have modified the original Cosine Coefficient Similarity Measurement to rank the candidate answers.

Keywords: *TF-IDF; Vector Space Model, Cosine Similarities, Term-Document, Term-Query Matrices, Dot Products.*

I. INTRODUCTION

Retrieval of relevant documents using given set of document collection and Query Document, consist of a keywords which convey the semantics of the information need, we need to retrieve relevant documents. The following major models have been developed to retrieve information: the Boolean Model, the Statistical Model, it's including the Vector Space and the Probabilistic Retrieval Model and the Linguistic Model and Knowledge-Based models. Both retrieval and browsing are, in the language of the World Wide Web, `pulling' actions. That is, the user requests the information in an interactive manner. An alternative is to do retrieval in an automatic and permanent fashion using software agents which push the information towards the user.

Major Information Retrieval Models

1. Standard Boolean
2. Narrowing and Broadening Techniques
3. Smart Boolean
4. Extended Boolean Models
5. Statistical Model
   (a).Vector Space Model
   (b).Probabilistic Model
   (c). Latent Semantic Indexing
6. Linguistic and Knowledge-based Approaches
   (a). DR-LINK Retrieval System

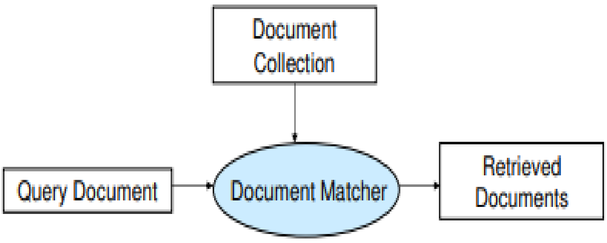


Figure 1: Basic Information Document Retrival System

The Boolean Model of information retrieval (BIR) is a classical information retrieval (IR) model and is the first and most adopted one, at the same time, the first and most-adopted one. It is used by many Information Retrieval systems used day. The Boolean Model of information retrieval (BIR) is a classical information retrieval (IR) model and, at the same time, the first and most-adopted one. It is used by many information retrieval (IR) systems to this day. The Boolean Information Retrieval is based on Boolean logic and classical set theory in that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval is based on whether or not the documents contain the query terms **[4].**Vector Space Model is an algebraic model its involving two steps: (i).We represent the text documents into vector of words and (ii).We transform to numerical format so that we can apply any text mining techniques.

As mentioned earlier, a Boolean query can be described in terms of the following four operations: degree and type of coordination, proximity constraints, field specifications and degree of stemming as expressed in terms of word/string specifications. Each of the four kinds of operations in the query formulation has particular operators, some of which tend to have a narrowing or broadening effect.

They have some of the disadvantages of the traditional Boolean , describe such a method, called Smart Boolean, that to help users construct and modify a Boolean query as well as make better choices along the four dimensions that characterize a Boolean query. Several methods have been developed to extend the Boolean model to the following issues: 1) The Boolean operators are too strict and ways need to be found to soften them. 2) The standard Boolean approach has no provision for ranking.

The vector space, probabilistic retrieval model and Latent Semantic Indexing & Clustering s are the three major examples of the statistical retrieval approach. Both models use statistical information in the form of term frequencies to determine the relevance of documents with respect to a query. The vector space model represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents. The probabilistic retrieval model is based on the Probability Ranking Principle, which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query. The two central quantities used are the inverse term frequency in a collection (idf), and the frequencies of a term i in a document j (*freq(i,j)*). In the probabilistic model, the weight computation also considers how often a term appears in the relevant and irrelevant documents.

In the simplest form of automatic text retrieval, users enter a string of keywords that are used to search the inverted indexes of the document keywords is called Linguistic and Knowledge-based Approaches. This approach retrieves documents based solely on the presence or absence of exact single word strings query. DR-LINK is based on the concept of that retrieval should take place at the conceptual level and not at the word level. The DR-LINK retrieval system represents content at the conceptual level rather than at the word level to reflect the multiple levels of human language comprehension. The text representation combines the lexical, syntactic, semantic, and discourse levels of understanding to predict the relevance of a document. DR-LINK accepts natural language statements, which it translates into a precise Boolean representation of the user's relevance requirements. It also produces a summary-level, semantic vector representations of queries and documents to provide a ranking of the documents.

II. RELATED WORK

Jitendra Nath Singh and Sanjay Kumar Dwivedi [32] Analysis of Vector Space Model in Information Retrieval. National Conference on Communication techniques with impack of next generation.Maron and Kuhns [34] in early 1960, described probabilistic indexing technique in a mechanized library system yielding probable relevance. After word in 1983, Salton and McGill wrote a book [35] which discusses thoroughly the three classic models in information retrieval namely, the boolean, the vector, and the probabilistic models.The book by van Rijsbergen [17] covers the discussion on three classic models and majority of the associated technology of retrieval system. Frakes and Baeza-Yates [36] edited the book on information retrieval which mainly deals with the data structures used in general information retrieval systems. Also, it includes the issue of relevance feedback as well as some query modification techniques [7] and Boolean operations and their implementations [8]. Verhoeff, Goffman, and Belzer [37] described the shortfall of boolean queries for information retrieval. The concept of using boolean formalism in other frameworks had been the great interest area of the researchers. Lee et al proposed a thesaurus-based boolean retrieval system for ranking Vector space model has been the most popular model in information retrieval among the research vicinity because of the research outcome in indexing, term value specification in automatic indexing carried out by Salton and his associates [14, 07]. Most of this research deals with experiments in automatic document processing and different term weighting approaches for automatic retrieval [12].

III. METHODOLOGIES AND TECHNOLOGIES IMPLEMENTED IN QA SYSTEM

Some of the updated methodologies and technologies which are introduced in the papers published in the last five years are summarized. Each of them is presented with the structure as the motivation, explanation and evaluation.

IV. EVALUATIONS

Since we have modified both the traditional term-based document search strategy by expanding the search queries and the CCSM algorithm to calculate the similarity of the retrieved documents from the web pages, we needed to design two evaluation schemes to test those two modified algorithms separately. Instead of having subjective surveys collecting the users' experiences and opinions of our system, we preferred to evaluate them with some scalable and objective evaluations. Only the figures can prove the improvements and the performances of those two modified algorithms.

V. EVALUATION ON DOCUMENT RETRIEVAL STRATEGY

To evaluate our modified and improved document retrieval strategy which expands the received questions in our QA system, there are two factors need to be considered. First of all, one of them is how many relevant documents our system is able to retrieve from the document collection based on a question. The purpose of designing the search query expansion which is about adding the web page snippets to the query is helping our system retrieve more semantically related documents from the collection. Therefore, we are expecting our system is able to retrieve more relevant documents with the modified document retrieval strategy than using the traditional strategy. Another factor to consider the improvement is the number of irrelevant documents our system searches back. Since the raise of the number of the retrieved documents is not avoidable in our strategy, we need to calculate how many irrelevant documents are also sent back to the users which will disturb them searching their ideal answers.

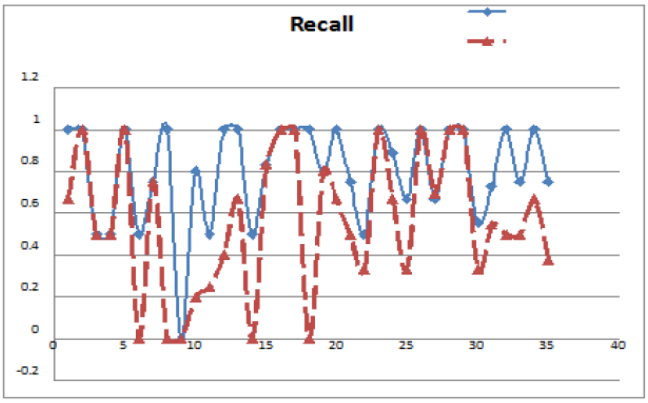


Figure 1

Figure 1 shown above is comprised by two groups of recall values. The dotted line with the triangle the triangle markers represents the 35 recall values produced by the traditional document retrieval strategy which only searches the documents with the original queries. The real line with the diamond markers denotes another 35 recalls which were calculated for our

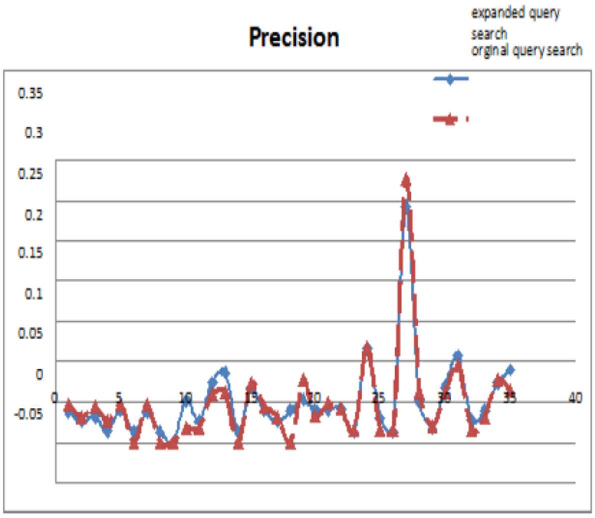
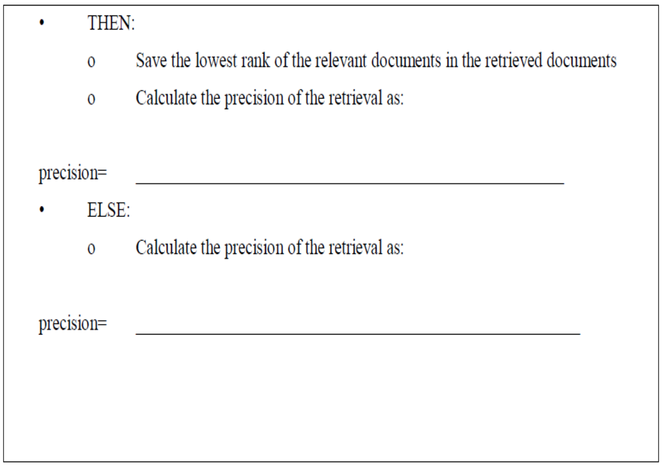


Figure 2

Comparing with the two averages of recalls and precisions, we found some unexpected results. Since the new retrieval strategy offered our system a 1.4 times higher average recall than the traditional strategy did, it means the new strategy helped our system retrieve more relevant documents. Based on the theory, the new strategy was also supposed to bring the system the side-effect that more irrelevant documents were retrieved at the same time. In other words, the new retrieval strategy was theoretically expected to perform an obvious lower precision value than the traditional strategy did. However, the evaluation results were in the contrary situation: the average precision of the new strategy was even slightly higher than the precision of the traditional strategy. It was because during the experiment, our system retrieved 0 relevant documents for some New Strategy of Calculating Precision.

- IF : all the relevant documents in the collection have been retrieved by the system



- THEN:
  - o Save the lowest rank of the relevant documents in the retrieved documents
  - o Calculate the precision of the retrieval as:

precision= ________________________________

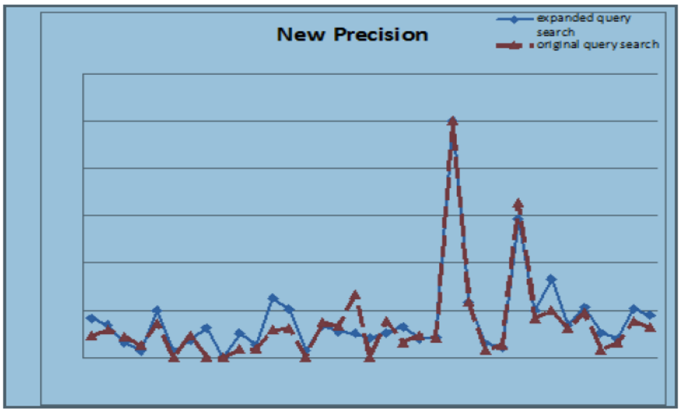
- ELSE:
  - o Calculate the precision of the retrieval as:

precision= ________________________________

Average of the new precisions produced by the traditional retrieval strategy was around 0.07. There was an obvious increase from the original average precisions to the new calculated average precisions. It was because when our system had successfully retrieved all of relevant documents in the collection for some questions; it also ranked them with the higher ranks. The users then did not need to go through the entire retrieved documents to read all the relevant documents. They stopped checking the documents at the lowest rank of the relevant documents. Thus, the new precision should be higher since the number of retrieved documents the users read was less than the total number of retrieved documents.

Based on the recalls and the new precisions we calculated while answering those 35 questions with the local document collection. Our new document retrieval strategy which expands the search queries with the snippets of their related web pages performed remarkably better retrieval results than the traditional document retrieval strategy did. The new strategy’s expected lower precision as the trade-off of its higher recall did not appear since it made sure the system retrieve at least parts of the relevant documents back to the users. On the contrary, the traditional document strategy made the system to be over precise during the retrieval that none of relevant documents were searched in some cases. Furthermore, as both of the document retrieval strategies is the term-based search strategies which means as long as a document shares at least one common term with the question it will be retrieved, the higher recalls and the lower precisions are expected 70. For the users who search a particular type of information such as the medical knowledge and the law cases in our system, the higher recall helps them find what they need eventually. The lower precision does not bother them much since they have to find the information they need no matter how long it takes.

Furthermore, we designed and conducted this evaluation to test the improvement of our new document retrieval strategy. During the ranking process, we did not use the MCCSM since the documents in the collection were not extracted from the related web pages. There were no corresponding web pages’ ranks to use to estimate the qualities of the local documents. We could not use the documents’ length as one of the factors to rank the retrieved documents since they did not use to be the paragraphs on some web pages. Thus, the evaluation results discussed above only describe the retrieval ability and improvement of our new document search strategy.



VI. EVALUATION ON OVERALL PERFORMANCE

Therefore, we have also designed another evaluation to test the whole performance of our QA system with the new retrieval and ranking strategies. There were two reasons we did not test the ranking algorithm MCCSM alone and compare the result with the CCSM’s. First of all, it is because of the two modifications made in the MCCSM. One of them is adding their corresponding web pages’ ranks into the similarity measurement to rank the retrieved documents. Another one is considering the length of each document during the ranking. As they were explained in the candidate answer ranking section, we treat those two new factors: web page rank and document length as two necessary.

VII. IMPLEMENTATION

In order to test and make use of the MCCSM algorithm and the modified and improved term-based document search strategy, we have successfully implemented them together as a complete and functioning QA system. This QA system is dynamically connected with the Google search engine to access the abundant and up-to-date online information. Users can ask their questions in natural English on the unlimited topics to our system. Based their questions, our system is able to answer them with numbers of related and ranked documents as the candidate answers. Thus, the users can find the needed and correct information among those retrieved documents with less time spent. The QA system then is able to improve the quality of the online information retrieval eventually.

During the implementation, we used Java with the NetBeans IDE 7.0.1 as our programming language and the coding environment. NetBeans offered us a visualized coding environment so that we could program and design the User Interface at the same time with a more straightforward view. The class-based and object-oriented features of Java programming language offered us an ideal programming structure to use to implement our algorithms and strategies. We were able to code the different steps of the process as the different classes in the separate objects. Thus, it is easier and clearer to monitor and test their functions and performance individually. It will be also easier to control the affections of the future modifications.

CONCLUSION

In this paper, we have introduced and explained our QA system which was designed for improving the retrieval of information on the Internet. Several modified and improved algorithms were embedded into this system in order to attain this goal. In this section, we are going to summary the functions and the performances of the main components in our system and also have an overall conclusion of this complete system. The initial motivation of our QA system was based on the web search engines' search process and results. The users submit their questions to those web search engines in order to search some useful answers. There is a growing discrepancy between the retrieval approach used by existing commercial retrieval systems and the approaches investigated and promoted by a large segment of the information retrieval research community. The former is based on the Boolean or Exact Matching retrieval model, whereas the latter ones subscribe to statistical and linguistic approaches, also referred to as the Partial Matching approaches.

References

[1] https://www.datasciencecentral.com/profiles/blogs/information-retrieval-document-search-using-vector-space-model-in.

[2] http://web.letras.up.pt/bhsmaia/EDV/apresentacoes/Bradzil_ReprDocs_InfRetr.pdf –Informa

[3] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_917.

[4] https://en.wikipedia.org/wiki/Boolean_model_of_information_retrieval.

[5] https://www.codeproject.com/Articles/375219/Boolean-Retrieval-Model[5].

[6] https://aspoerri.comminfo.rutgers.edu/InfoCrystal/Ch_2.html.

[7] Jitendra Nath Singh and Sanjay Kumar Dwivedi, "Analysis of Vector Space Model in Information Retrieval", National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC 2012, Proceedings published by International Journal of Computer Applications® (IJCA).

[8] Gaurav Batra, Mansi Goel. An improved answer retrieval system taping the linkage structure for noisy SMS queries. International Journal of Computer Applications, 2012.

[9] Ryuichiro Higashinaka, Hideki Isozaki. Corpus-based question answering for why-questions. International Joint Conference on Natural Language Processing, 2008.

[10] Richard C. Wang, Nico Schlaefer, WilliamW. Cohen, Eric Nyberg. Automatic set expansion for list Question Answering. EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing Pages 947-954, 2008.

[11] João Silva, Luísa Coheur, Ana Cristina Mendes, AndreasWichert. From symbolic to sub- symbolic information in question classification. Artificial Intelligence Review, Volume 35 Issue 2, Pages 137-154, February 2011.

[12] Cheng-Lung Sung, Cheng-Wei Lee, Hsu-Chun Yen, Wen-Lian Hsu. An alignment-based surface pattern for a Question Answering system. Integrated Computer-Aided Engineering - Selected papers from the IEEE Conference on Information Reuse and Integration (IRI), July 13-15, 2008, Volume 16 Issue 3, Pages 259-269, August 2009.

[13] Tianyong Hao, Dawei Hu1, Liu Wenyin and Qingtian Zeng. Semantic patterns for user- interactive question answering. Semantics, Knowledge and Grid, 2006. SKG '06. Second International Conference, 2006.

[14] Leila Kosseim , Jamileh Yousefi. Improving the performance of Question Answering with semantically equivalent answer patterns. Journal: Data & Knowledge Engineering, Volume 66 Issue 1, Pages 53-67, July, 2008.

[15] Saeedeh Momtazi, Dietrich Klakow. A word clustering approach for language model-based sentence retrieval in Question Answering systems. CIKM '09 Proceedings of the 18th ACM Conference on Information and Knowledge Management, Pages 1911-1914, 2009.

[16] S.Kalaivani, K. Duraiswamy. Methodology for converting question to query form in Question Answering for automatic learning system. European Journal of Scientific Research, 2012.

[17] Gulfishan Firdose Ahmed, Nepal Barskar. An Approach for extracting exact answers to Question Answering (QA) system for english sentences. International Conference on Communication Technology and System Design, 2011.

[18] D.S. Wang. A domain-specific Question Answering system based on ontology and question templates. Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD), 11th ACIS International Conference, 2010.

[19] Shiyan Ou, Viktor Pekar, Constantin Orasan, Christian Spurk, Matteo Negri. Development and alignment of a domain-specific ontology for Question Answering. In Proceedings of the 6th Edition of the Language Resources and Evaluation Conference (LREC-08), 2008.

[20] Óscar Ferrández, Rubén Izquierdo, Sergio Ferrández, José Luis Vicedo. Addressing ontology-based Question Answering with collections of user queries. Lexical and Syntactic Knowledge For Information Retrieval, 2011.

[21] Hyo-Jung Oh, Sung Hyon Myaeng, and Myung-Gil Jang. Enhancing performance with a learnable strategy

for Multiple Question Answering Modules. ETRI Journal, Volume 31, Page 419-428, August, 2009.

[22] Antonio Ferrández, Jesús Peral. The benefits of the interaction between data warehouses and Question Answering. EDBT '10 Proceedings of the 2010 EDBT/ICDT Workshops, Article No. 15, 2010.

[23] V. Rieser, O. Lemon. Does this list contain what you were searching for learning adaptive dialogue strategies for Interactive Question Answering. Natural Language Engineering archive, Volume 15 Issue 1, Pages 55-72, January 2009.

[24] Sebastian Varges, Fuliang Weng, Heather Pon-Barry. Interactive Question Answering and constraint relaxation in spoken dialogue system. Natural Language Engineering, Volume 15 Issue 1, Pages 9-30, January 2009.

[25] Wael Salloum. A Question Answering system based on Conceptual Graph Formalism. KAM '09 Proceedings of the 2009 Second International Symposium on Knowledge Acquisition and Modeling, Volume 03, Pages 383-386, 2009.

[26] Liang Zhenqiu. Design of automatic Question Answering system Base on CBR. 2012 International Workshop on Information and Electronics Engineering, 2012.

[27] Ulli Waltinger, Alexa Breuing, IpkeWachsmuth. Interfacing virtual agents with collaborative knowledge open domain Question Answering using wikipedia-based topic models. In proceeding of IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011.

[28] Detle Koll, Thomas Polzin. Providing computable guidance to releevant evidence in Question-Ansering system. Application Number: 13/025,051, Publication Number: US 2012/0041950 A1, Filing Date: Feb 10, 2011.

[29] Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, Philipp Cimiano. Template-based Question Answering over RDF Data. WWW '12 Proceedings of the 21st International Conference on World Wide Web, Pages 639-648, 2012.

[30] Elif Aktolga, James Allan, David A. Smith. Passage reranking for Question Answering using syntactic structures and answer types. ECIR'11 Proceedings of the 33rd European Conference on Advances in Information Retrieval, Pages 617-628, 2011.

[31] Arnaud Grappy, Brigitte Grau. Answer type validation in Question Answering Systems. RIAO '10 Adaptivity, Personalization and Fusion of Heterogeneous Information, Pages 9-15, 2010.

[32] Jitendra Nath Singh and Sanjay Kumar Dwivedi. Analysis of Vector Space Model in Information Retrieval. National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC 2012 Proceedings published by International Journal of Computer Applications® (IJCA).

[33] A. B. Manwar, Hemant S. Mahalle, K. D. Chinchkhede and Dr. Vinay Chavan A Vector Space Model for Information Retrieval: A MATLAB APPROACH. Indian Journal of Computer Science and Engineering (IJCSE).

[34] Maron, M.E. and Kuhns, J.L. (1960) On Relevance, Probabilistic Indexing and Information Retrieval. Journal of the ACM (JACM), 7, 216-244.

[35] Salton, G. and McGill, M.J. (1983) Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York.

[36] William B. Frakes, Software Engineering Guild, Sterling, VA, USA Ricardo Baeza-Yates, Universidad de Chile Information Retrieval: Data Structures and Algorithms

[37] Jaime Chon Ryan Roberts Implementation of Query option in data service.