An introduction to modern Information retrieval: A brief overview

G Selva Priya[1], T.Primya[2], V.Suresh[3], R.Ashwini[4] , G.kanagaraj[5]
[1, 2, 3, 4, 5] Department of Computer Science and Engineering
[1, 2, 3] Dr.N.G.P Institute of Technology
[4] IFET Engineering and Technology
[5] Kumaraguru college of Engineering

Abstract- *For thousands of years people have realized the importance of archiving and finding information. With the advent of computers, it became possible to store large amounts of information; and finding useful information from such collections became a necessity. The field of Information Retrieval (IR) was born in the 1950s out of this necessity. Over the last forty years, the field has matured considerably. Several IR systems are used on an everyday basis by a wide variety of users. This article is a brief overview of the key advances in the field of Information Retrieval, and a description of where the state-of-the-art is at in the field.*

Keywords- Information Retrieval, archiving, advent, state-of-the-art.

I. INTRODUCTION

This article guides a stepwise walkthrough for Information retrieval (IR) which is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

As defined in this way, information retrieval used to be an activity that only a few people engaged in: reference librarians, paralegals, and similar professional searchers. Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email.

II. CATEGORIES OF IR

Information retrieval can be divided into three types such as unstructured data, structured data, semi structured data.

A. Unstructured data

The term "unstructured data" refers to data which does not have clear, semantically overt, easy-for-a-computer structure. It is the opposite of structured data, the canonical example of which is a relational database, of the sort companies usually use to maintain product inventories and personnel records. In reality, almost no data are truly "unstructured". This is definitely true of all text data if you count the latent linguistic structure of human languages. But even accepting that the intended notion of structure is overt structure, most text has structure, such as headings and paragraphs and footnotes, which is commonly represented in documents by explicit markup (such as the coding underlying webpages).

B. semi structured data

IR is also used to facilitate "semistructured" search such as finding a document where the title contains Java and the body contains threading. The field of information retrieval also covers supporting users in browsing or filtering document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. It is similar to arranging books on a bookshelf according to their topic. Given a set of topics, standing information needs, or other categories (such as suitability of texts for different age groups), classification is the task of deciding which class(es), if any, each of a set of documents belongs to. It is often approached by first manually classifying some documents and then hoping to be able to classify new documents automatically.

C. Structured data

Structured data refers to information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations; whereas unstructured data is essentially the opposite. The lack of structure makes compilation a time and energy-consuming task. It would be beneficial to a company across all business strata to find a mechanism of data analysis to reduce the costs unstructured data adds to the organization.

III. AN EXAMPLE INFORMATION RETRIEVAL PROBLEM

But for many purposes, you do need more:

1. To process large document collections quickly. The amount of online data has grown at least as quickly as the speed of computers, and we would now like to be able to search collections that total in the order of billions to trillions of words.

2. To allow more flexible matching operations. For example, it is impractical to perform the query Romans NEAR countrymen with grip, where NEAR might be defined as "within 5 words" or "within the same sentence".

3. To allow ranked retrieval: in many cases you want the best answer to an information need among many documents that contain certain words.

IV. SUMMING UP

The field of information retrieval has come a long way in the last forty years, and has enabled easier and faster information discovery. In the early years there were many doubts raised regarding the simple statistical techniques used in the field. However, for the task of finding information, these statistical techniques have indeed proven to be the most effective ones so far. Techniques developed in the field have been used in many other areas and have yielded many new technologies which are used by people on an everyday basis, e.g., web search 7 engines, junk-email filters, news clipping services. Going forward, the field is attacking many critical problems that users face in today's information-ridden world. With exponential growth in the amount of information available, information retrieval will play an increasingly important role in future.

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, pages 194–218, 1998.

[2] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35(12):29–38, 1992.

[3] Chris Buckley, James Allan, Gerard Salton, and Amit Singhal. Automatic query expansion using SMART: TREC 3. In Proceedings of the Third Text REtrieval Conference (TREC-3), pages 69–80. NIST Special Publication 500-225, April 1995.

[4] Chris Buckley, Gerard Salton, and James Allan. Automatic retrieval with locality information using SMART. In Proceedings of the First Text REtrieval Conference (TREC-1), pages 59–72. NIST Special Publication 500-207, March 1993.

[5] Vannevar Bush. As We May Think. Atlantic Monthly, 176:101–108, July 1945.

[6] C. W. Cleverdon. The Cranfield tests on index language devices. Aslib Proceedings, 19:173–192, 1967.

[7] W. B. Croft and D. J. Harper. Using probabilistic models on document retrieval without relevance information. Journal of Documentation, 35:285–295, 1979.

[8] J. L. Fagan. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. Journal of the American Society for Information Science, 40(2):115–139, 1989.

[9] G. Grefenstette, editor. Cross-Language Information Retrieval. Kluwer Academic Publishers, 1998.

[10] A. Griffiths, H. C. Luckhurst, and P. Willett. Using interdocument similarity in document retrieval systems. Journal of the American Society for Information Science, 37:3–11, 1986.

[11] D. K. Harman. Overview of the first Text REtrieval Conference (TREC-1). In Proceedings of the First Text REtrieval Conference (TREC-1), pages 1–20. NIST Special Publication 500-207, March 1993.

[12] David Hull. Stemming algorithms - a case study for detailed evaluation. Journal of the American Society for Information Science, 47(1):70–84, 1996.

[13] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. Retrieving spoken documents by combining multiple index sources. In Proceedings of ACM SIGIR'96, pages 30–38, 1996.

[14] K. Sparck Jones. Automatic Keyword Classification for Information Retrieval. Butterworths, London, 1971.

[15] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28:11–21, 1972.

[16] K. Sparck Jones and P. Willett, editors. Readings in Information Retrieval. Morgan Kaufmann, 1997.

[17] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1957.

[18] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. Journal of the ACM, 7:216–244, 1960.

[19] Marius Pasca and Sanda Harabagiu. High performance question/answering. In Proceedings of the 24th International Conference on Research and Development in Information Retrieval, pages 366–374, 2001.

[20] S. E. Robertson. The probabilistic ranking principle in IR. Journal of Documentation, 33:294–304, 1977.

[21] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3):129–146, May-June 1976.

[22] S. E. Robertson and S. Walker. Some simple effective approximations to the 2–poisson model for probabilistic weighted retrieval. In Proceedings of ACM SIGIR'94, pages 232–241, 1994.