Information Noise and the Role of Ontology in Information Retrieval System

Dr. S. K. Panday

Deputy Librarian, Central Library, Kumaun university, Nainital Ph-(o) 05942-232294, (M) 9412986130

Abstract Recent Information Retrieval systems have been developed according to the time of web 1.0, when the web pages were very few and it was easy to filter them but now in the time of information explosion, when everyone contributing on web these Information Retrieval system starting to generate Information Noise; they are retrieving information which is not according to the need of the user. Ontology which is one of the part of Semantic web provides richer integration and interoperability of data and permit the development of application that search across diverse area of information or merge information to reduce information noise. Present paper is the study of traditional Information Retrieval System their mechanism and how Ontologies can be efficiently applied to reduce Information noise by providing semantic representation of information in documents.

Keywords: Information Retrieval, Ontology, Semantic Web, Information noise

----------♦----------

INTRODUCTION

Irrelevant Information is a well-recognized problem now days. Earlier was the problem of non-availability of information but now we face the problem of huge information, majority of which is according to our query but not according to our need. The right information/required information/efficient information/effective information from Information retrieval system /Internet/ Intranet has become an imminent issue these days. There are various indexing system, cataloguing system and search engines available on internet but methodology/ technology of these system to retrieve information is limited and not helpful these days.

There are already exists several well-known problem for traditional information retrieval system, for instance, the vocabulary inconsistency between user queries and information actually provided[1] and the simple keyword-matching approach statistically flavored in the sense of exploiting frequency data about the occurrences and co-occurrence of natural language terms [2, 3, 4].

This paper is prepared in three sections; Section 01 briefly describe the issue of information noise, section 02 provides the general introduction about Information Retrieval system and its components, section 03 tells about ontology and how it is helpful for reducing information noise.

INFORMATION NEED

Information need refers to the type of information sought by the user. Belkin et. al. define an information need as a problematic situation where a person cannot attain some goals due to inadequacy of resources or knowledge [20,24]. Kuhlthau defines an information need as the gap between the user's problem or topic and what the user needs to know to solve a problem [21, 24].

Information needs have been classified in various manners by different researchers. Tague-Sutcliffe [22, 24] classified information needs into categories such as quick reference questions, how-to-do questions, questions that involve collecting and synthesizing information about a topic, and doing a literature search for a project. These were based on the kind of information required for the user task or question for which information is sought, as well as whether there would be variation among users about expected results. Glover et. al. [23, 24] suggested categories based on the kind of information sought. Categories include research papers, home pages of research organizations, topical current events, and introductory articles.

INFORMATION NOISE

Each of the search process outputs generates information noise due to the inherent characteristics of the process itself [24]. A non-ideal query, simple keyword matching approach, common words search formulation, may generate irrelevant search results that are hard to filter. As

a result, we can say that the particulars of the query act as a source of the information noise.

Keyword based search engine generate information noise when the query is on compound or complex subject, For example if the query is given by the user like “famous hotels in Newyork” the search engine will start its searching for the document which contains all the above 4 words(as they are is quotes, otherwise it will search only three words) But the results are (i) A story that contains the word “one day in a famous hotel that resides in Newyork” (ii) A list of document that contain only the name of the hotels without specifying any address (iii) A list of documents with the keyword “famous” and “hotels” without including the city “Newyork”. (iv)Document which specifies the special features of the city Newyork and so on. Most of the time the search engine will retrieve the documents that are not relevant to the user query. [5] and through this, search engine generate information noise.

INFORMATION RETRIEVAL SYSTEM

Information retrieval system finds answers and information that already exist in a system. Information retrieval system Search by query (as in a search engine) and often deal with whole document, such as books and journal and it is not search by navigation which is following links, as in subject directory. It is used for Database Management system [6]

Development of Information Retrieval System [5]:

First Generation: (Keyword Based)

- Documents are retrieved based on the ranking of web pages that has the maximum number of query term.
- Every page was already indexed based upon keyword Second Generation: (Ranking based) Ranking based upon

- Keyword Focused Anchor Text from External Links
- External Link Popularity
- Diversity of Link Source
- Usage of Keyword in the Title Tag
- Trustworthiness of the Domain [A conceptual Framework]

Third Generation (Semantic Based) (XML, RDF, OWL) XML

- User defined Tags
- Lagging of Semantic

RDF

- Semantics are added
- Represented as triples(Resource, Property, Value)

OWL

- Scope of Properties
- Disjointness of classes
- Cardinality restrictions
- Boolean combination of classes
- Special characteristics of properties

How Information retrieval system works:

STEPS IN THE INFORMATION RETRIEVAL PROCESS

1. Indexing: Indexing means to indicate where a particular document of information is reside through an artificial language, In Information retrieval system indexing is done on a database system. Indexing is a process (manual or automated) of making statements about a document, lesson, person and so on, in accordance with the conceptual schema.

i. Manual Indexing

Indexing can be document-oriented (the indexer capture what the document is about) or request-oriented (the indexer assesses the document’s relevance to subjects and other features of interest to users)

ii. Automatic Indexing

Automatic indexing begins with raw feature extraction, such as extracting all the words from a text, followed by refinements, such as eliminating stop words(and, it, of),stemming (pipes=pipe), counting (using only the most frequent words), and mapping to concepts using a thesaurus (tube and pipe map to same concept). A program can analyze sentence structure to extract phrases, for images; extractable feature include color distribution or shapes. For music, extractable features include frequency of occurrence of notes or chords, rhythm and melodies.

2. Query formulation

Retrieval means using the available evidence to predict the degree to which a document is relevant or useful for a given user need as described in a free-form query description, also called topic description or query statement. The query description is transformed, manually or automatically, into a formal query representation (also called query formulation or query for short) that combines features that predict a document's usefulness. The query expresses the information need in terms of the system's conceptual schema, ready to be matched with document representations. A query can specify text words or phrases the system should look for (free-text search) or any other entity feature, such as descriptors assigned from a controlled vocabulary, an author's organization, or the title of the journal where a document was published.

A query can simply give features in an unstructured list (for example, a "bag of words") or combine features using Boolean operators (structured query)

3. Matching the query representation with entity representations

The match uses the features specified in the query to predict document relevance. In exact match the system finds the documents that fill all the conditions of a Boolean query (it predicts relevance as 1 or 0). To enhance recall, the system can use synonym expansion (if the query asks for pipe, it finds tubes as well) and hierarchic expansions or inclusive searching (it finds capillary as well). Since relevance or usefulness is a matter of degree, many IR systems (including most Web search engines) rank the results by a score of expected relevance (ranked retrieval)[6]

PROBLEM WITH INFORMATION RETRIEVAL SYSTEM:

According to Sparck Jones [7, 4] following are the problem with Information retrieval system.

1. Knowledge Representation. IR's representation of entities and relation is very weak, "Concept names are not normalized, and description are mere sets of independent terms without structure….. Concepts and topics, term and description meanings are left implicit…. The relation between terms is only association based on co-presence….."

2. Reasoning. The weak reasoning in IR is "looking at what is in common between descriptions and preferring one item over another because more in shared (whether as different words or, via weighting, occurrences of the same word)…. The probabilistic network approach, that allows for more varied forms of search statement and matching condition, does not alter the basic style of reasoning."

3. Learning. Loosely speaking, the relevance feedback of IR can be considered as forms of learning. [8]

The traditional statistical model-based Information Retrieval system was successful in past but now it is facing a tuff task to full fill the users need, it is the need of the hour to bring an IR or change in the components so that it will generate less information noise.

ONTOLOGY

According to Oxford English dictionary Ontology is " the science or study of being" In Artificial Intelligence it is usually attribute the notion of ontology to, essentially, the specification of a conceptualization-that is, defined terms and relationships between them, usually in some formal and preferably machine-readable manner.[9]

Ontology can be defined as a set of knowledge terms, including the vocabulary, the semantic interconnections, and some simple rules of inference and logic for some particular topic. For example, the ontology of cooking and cookbooks includes ingredients, hot to stir and combine the ingredients, the difference between simmering and deep-frying, the expectation that the products will be eaten or drunk, that oil is for cooking or consuming and not for lubrication, and so forth.

as we understand that information noise is generated because of wrong indexing and possibly wrong cataloguing or limited cataloguing, in case of automated indexing system which is Key word indexing based that left the relational words is, am, with etc which is very important from user point of view.

HOW ONTOLOGY IS HELPFUL IN LIMITING THE INFORMATION NOISE

One of the reasons for why IR systems do not have an explicitly defined domain of interest to the user is that most users tend to use very few terms (3 or less) in their search queries [11, 12, 17]. As a result, the systems cannot understand the context of the user's query, which results in lower precision. By adding more relevant terms to the query, the domain of interest can, to some extent, be identified. However, adding both correct and distinctive terms is not always trivial, since the user needs knowledge about the terminology used in that particular domain to find those correct terms. A novel and promising approach is concept-based search [13, 14, 15, 17]. With this approach, the burden of knowing how the documents are written is taken off by the user and hence the user can focus on

searching on a conceptual level instead. One problem with this approach is to find good concepts.

Concepts and, in particular, relations between them can be specified in ontologies. Ontologies define concepts and the relationships among them [16,17]; therefore, they are often used to capture knowledge about domains. A growing number of IR systems make use of ontologies to help clarifying the information needs of the users; however, a concern with these semantic approaches is the integration with traditional commercial search technologies.

How Ontology is formulated [18]

Step-1: Determine Scope

To determine the scope of the ontology, we have to answer several basic questions:

- What is the domain that the ontology will cover?
- For what we are going to use the ontology?
- For what types of questions the information in the ontology should provide answers?
- Who will use the ontology?

The answers to these questions may change during the ontology-design process, but at any given time they help limit the scope of the model.

Step-2: Enumerate important terms in the ontology

It is useful to write down a list of all terms we would like either to make statements about or to explain to a user.

- What are the terms we would like to talk about?
- What properties do those terms have?
- What would we like to say about those terms?

Step-3: Define the classes and the class hierarchy

There are several possible approaches in developing a class hierarchy [19]

- A top-down development process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts.
- A bottom-up development process starts with the definition of the most specific classes, the leaves of the hierarchy, with subsequent grouping of these classes into more general concepts.
- A combination development process is a combination of the top-down and bottom-up approaches: We define the more salient concepts first and then generalize and specialize them appropriately.

Step-4: Define the properties of classes—slots

The classes alone will not provide enough information to answer the competency questions from Step-1. Once we have defined some of the classes, we must describe the internal structure of concepts.

Step-5: Define the facets or constraints of the slots

Slots can have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take.

- **Slot cardinality**

Slot cardinality defines how many values a slot can have. Some systems distinguish only between single cardinality (allowing at most one value)and multiple cardinality (allowing any number of values).

Some systems allow specification of a minimum and maximum cardinality to describe the number of slot values more precisely. Minimum cardinality of N means that a slot must have at least N values. Maximum cardinality of M means that a slot can have at most M values.

- **Slot-value type**

A value-type facet describes what types of values can fill in the slot. Here is a list of the more common value types:

- String is the simplest value type which is used for slots such as name: the value is a simple string
- Number (sometimes more specific value types of Float and Integer are used) describes slots with numeric values.
- Boolean slots are simple yes–no flags.
- Enumerated slots specify a list of specific allowed values for the slot.
- Instance-type slots allow definition of relationships between individuals. Slots with value type Instance must also define a list of allowed classes from which the instances can come. [3]
- **Domain and range of a slot**

Allowed classes for slots of type Instance are often called arrange of a slot. allow restricting the range of a slot when the slot is attached for a particular class.

Step-6: Create instances

The last step is creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires (1) choosing a class, (2) creating an individual instance of that class, and (3) filling in the slot values.

CONCLUSION:

Due to information explosion Information Retrieval systems are facing challenges in providing required information in a time frame with authenticity. This leads to a situation where information is present but users are not in a position to retrieve it with authenticity in a effective and efficient manner. In near future ontology is going to play major role in retrieval systems and this will lead us to situation where information explosion or noise can't hamper information retrieval or information use.

REFERENCE

1. Bates, M.J. (1998) Indexing and access for digital libraries and the Internet: Human, database, and domain factors, Journal of the American Society for Information Science, 49(13), 1998,1185-1205.

2. Sparck-Jones, K., & Willett, P. (1997) Reading in information retrieval. Morgand Kaufmann, 1997.

3. Ding, Y., Chowdhury, G.G., Foo, S. (2000)Incoporating the results of co-word analyses to increase search variety for information retrieval. Journal of Information Science, 26 (6), 2000, 429-452.

4. Ding, Y. IR and AI: The role of ontology

5. Maheswari, J, Karpagam, G.R. (2010) A Conceptual Framwork for Ontology based Information retrieval. International Journal of Engineering Science and Technology. Vol.2(10),2010,5679-5688.

6. www.dsoergel.com/NewPublications/HCIEncyclopediaIRShortEForDS.pdf

7. Sparck, Jones, K.(1999) Information retrieval and artificial intelligence. Artificial Intelligence, 114, 1999, 257-281

8. Michalski, R & Kaufmann, K.(1997) Data mining and knowledge discovery: A review of issues and multi-strategy, approach. In Machine Learning and Data Mining Methods and Applications, Jhon wiley & Sons Ltd. 1997.

9. Gruber, T.R.(1993) A Translation Approach to Portable Ontologies. Knowledge Acquisition, 5(2), 1993, 199-220.

10. Hendler, James (2001) Agents and the Semantic Web: The Semantic Web. IEEE Intelligent Sustems, 16(2), 2001, 30-37.

11. Gulla, J.A., Auran, P.G., Risvik, K.M.(2002) Linguistic Techniques in Large-Scale Search Engines. Fast Search & Transfer ASA, Springer-verlag, Berlin, LNCH 2553, 2002, 218-222.

12. Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.(2001) Searching the Web: the public and their queries. J. Am. Soc. Inf. Sci. Technol. 52, 2001, 226-234

13. Grootjen, F.A., van der Weide, T.P.(2006) Conceptual query expansion. Data & Knowledge Engineering. 56, 2006, 174-193.

14. Qiu, Y., Frei, H.-P.(1993) Concept based query expansion. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Pittsburgh Pennsylvania, USA. 1993, 160-169.

15. Chang, Y., Ounis, I., Kim, M. (2006) Query reformulation using automatically generated query concepts from a document space. Information Processing and Management. 42, 2006, 453-468.

16. Gruber, T.R.(1993) A translation approach to portable ontology specifications. Knowledge Acquisition. 5, 1993, 199-220.

17. Stein L. Tomassen.(2006) Research on Ontology-Driven Information Retrieval OTM Workshops 2006. LNCS 4278, 2006, 1460 – 1468.

18. Natalya F. Noy and Deborah L. McGuinness Ontology Development 101: A Guide to Creating Your First Ontology, Stanford University, Stanford, CA, 94305.

19. Uschold, M. and Gruninger, M. (1996) Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 11(2), 2006.

20. Belkin, N. et. Al (1992). Information filtering and information retrieval: two sides of the same coin? Communications of the ACM. 35(12), 1992, 29 – 38.

21. Kuhlthau, C. C. (1991) Inside the Search Process: Information Seeking from the User's Perspective. Journal of the American Society for Information Science, 42 (5), 1991, 361-371.

22. Tague-Sutcliffe, J. (1992) Measuring the informativeness of a retrieval process, Proceedings of the 15th annual International ACM SIGIR Conference on Research and development in information retrieval, June 1992.

23. Gulli, A., and Signorini, A.(2005) Posters: The indexable web is more than 11.5 billion pages, Special interest tracks and posters of the 14th International Conference on World Wide Web, May 2005.

24. Shailja Venkatsubramanyan, and Steph En K Kwansusing. (2008) Information Noise to compute the Economic Benefit of a search service, Journal of Information Technology Management. XIX(1), 2008.