Information Retrieval Strategies

Presented by

Patrick Hoey

Information Retrieval Strategies For Keyword-Based Searches

- Information is stored in various types of media, ranging from customer account databases to CD-ROM encyclopedias to the internet
- The challenge lies in how to effectively retrieve the correct set of data, with the relevant information on it for the user

Information Retrieval Strategies For Keyword-Based Searches (Cont.)

- For databases, the information is stored in various tables which define the type of information which the data should represent. To retrieve the data, the user would type in a query, which is a formal request for particular information from the database.
- An example of a standard SQL query would be:

```
SELECT * FROM address WHERE town='Lowell' ;
```

Information Retrieval Strategies For Keyword-Based Searches (Cont.)

• Database techniques are not suitable for web queries though because the number of available documents is too high and there are too many users simultaneously accessing the system. In addition, web pages that contain multimedia occupy more memory space.

• The web cannot be organized, stored and indexed like a database. In addition the best query format is not known.

Information Retrieval Strategies For Keyword-Based Searches (Cont.)

• Indexing is used in a different way on the web. It involves building a data structure that facilitates searching. Often it involves finding index terms which summarize the information contained in the documents. The methods of indexing include manual indexing, automatic indexing, intelligent agent-based indexing and annotation based indexing. An indexing strategy must take into account coverage, typos, non-text content, redundancy and frequent updates.

Information Retrieval Strategies For Keyword-Based Searches (Cont.)

- Due to the fact that the internet is indexed in this fashion, the trend is to use keyword based search engines to find the relevant information on the web
- The two main metric techniques used to gauge the effectiveness of a search after a query are recall and precision.
- Recall is the proportion of relevant items in the entire repository which have been retrieved
- Precision is the proportion of retrieved items which are relevant.

Information Retrieval Strategies For Keyword-Based Searches (Cont.)

- Browsing is a technique used to aid in the iterative and ill-defined nature of information seeking, but leads to loss of direction and overly narrow searches
- Keyword based searches by themselves tend to have poor recall and low precision
- A structured query provides a means to direct the search, but often relies on the users understanding of a complex query language and proper vocabulary to be effective
- The following techniques should help in information retrieval of documents on the internet using standard web based search engines

Information Retrieval Strategies For Keyword-Based Searches (Cont.)

1. Analyze your topic to decide where to begin
2. Use formatting techniques on your query
3. Learn as you go and vary your approach based on the information retrieved

1. Analyze Your Topic To Decide Where To Begin

- **Does your topic have distinctive words or phrases?**

  i.e., methernitha, unique meaning

  i.e., "affirmative action", specific, accepted meaning in word cluster

- **Does your topic have NO distinctive words or phrases you can think of? You have only common or general terms that get the "wrong" pages.**

  i.e., "order out of chaos", used in too many contexts to be useful

  i.e., sundiata, retrieves a myth, a rock group, a person,

1. Analyze Your Topic To Decide Where To Begin (Cont.)

- **Does your topic seek an overview or cover a broad topic?**

  i.e., Victorian literature, alternative energy sources, concepts generally well accepted as specific

- **Does your topic specify a narrow aspect of a broad or common topic?**

  i.e., automobile recyclability, want current research, future designs, not how to recycle or oil recycling or other community efforts

1. Analyze Your Topic To Decide Where To Begin (Cont.)

- **Does your topic have synonymous, equivalent terms, or variant spellings or endings that need to be included?**

  i.e., echinoderm OR echinoidea OR "sea urchin", any may be in useful pages

  i.e., "cold fusion energy" OR "hydrogen energy", some use one term, some the other -- need to find both although not precisely interchangeable or equivalent

  i.e., millennium millennial millenium millenial "year 2000" "year 1000", etc., Pages you want may contain any or all.

2. Use Formatting Techniques On Your Query

Are you looking for a **proper name** or a **distinct phrase** ?

- The name of an organization or society or movement
- A proper name or an individual
- A distinctive string of words generally associated with your topic

Can you think of an organization, proper name, or phrase to search for? It might help zoom in on the pages you want.

2. Use Formatting Techniques On Your Query (Cont.)

PHRASE SEARCHING is a feature you want in every search tool you choose.

Requires your terms all to appear in exactly the order you enter them. Enclose the phrase in double quotations " "

Examples:

"affirmative action"

"world health organization"

"a person's name"

In Infoseek, capitalizing initial letters will cause the terms to be searched as a phrase: **World Health Organization**

2. Use Formatting Techniques On Your Query (Cont.)

Are some of your terms **common words** with **many meanings and contexts** ?

- *Children* in conjunction with *television* and also *violence*
- *Censorship* as an aspect of *ethics* in *journalism*

2. Use Formatting Techniques On Your Query (Cont.)

BOOLEAN AND will help:

children AND television AND violence

journalism AND ethics AND censorship

+REQUIRES forces all the terms to be present in all documents retrieved, and is nearly equivalent to Boolean AND:

+children +television +violence

+journalism +ethics +censorship

2. Use Formatting Techniques On Your Query (Cont.)

SUB-SEARCHING may also be helpful.

After submitting the search **journalism ethics** , sub-search (at bottom results, search "only within these results" on **censorship** . Sub-search within these results for **NEA "National Endowment for the Arts"** or other aspects of the arts to further focus.

2. Use Formatting Techniques On Your Query (Cont.)

Do you anticipate lots of search **results with terms you do not want** ?

• Your search for *biomedical engineering and cancer* brings you lots of academic programs, and you want research reports. So you try to exclude documents continuing *Department of* or *School of*

2. Use Formatting Techniques On Your Query (Cont.)

BOOLEAN AND NOT will help:

"biomedical engineering" AND cancer AND NOT "Department of" AND NOT "School of"

or its **-EXCLUDES** near equivalent:

+"biomedical engineering" +cancer -"Department of" -"School of"

2. Use Formatting Techniques On Your Query (Cont.)

Are there **synonyms** , **spelling variations** , or **foreign spellings** for some of your terms?

- *women, females* with *networking*
- *Sarajevo, Sarayevo* with *peace*
- *literature, litterature* with *French, francaise*

2. Use Formatting Techniques On Your Query (Cont.)

BOOLEAN OR will help:

(women OR females) AND networking

(Sarajevo OR Sarayevo) AND peace

(literature OR litterature) AND (French or francaise)

2. Use Formatting Techniques On Your Query (Cont.)

Its equivalent **is specifying niether +Requires/-Excludes**

+networking women females

+peace Sarajevo Sarayevo

literature litterature +French

literature litterature +francaise

With +Require/-Excludes, OR is the default when you specify neither + to require nor - to exclude.

2. Use Formatting Techniques On Your Query (Cont.)

Are you looking for **home pages** and/or other documents **primarily about** your term(s)?

- The home page of the *American Dietetic Association*
- Pages primarily about *Affirmative Action*

2. Use Formatting Techniques On Your Query (Cont.)

LIMIT to TITLE FIELD IN DOCUMENTS

```
title:"American Dietetic Association"

title:"affirmative action"
```

2. Use Formatting Techniques On Your Query (Cont.)

Are you looking for terms with **many possible endings** ?

- *Feminism, feminist, feminine*
- *Children, child*

2. Use Formatting Techniques On Your Query (Cont.)

TRUNCATION permits retrieving all these variations in one search term:

femini* matches *feminine, feminist, feminism*, etc.

child* retrieves *child* and *children*

Some systems search word ending variants automatically. See the specific instructions for each of the recommended search tools

3. Learn As You Go And Vary Your Approach Based On The Information Retrieved

- Learn as you go by refining your search based on results
- Vary your approach based on the information retrieved by repeating the search on similar search terms and their combinations; try this on different search tools

Search Engines To Try

http://www.google.com/

http://www.altavista.com/

http://infoseek.go.com/

http://www.lycos.com/

http://www.yahoo.com/

http://www.astalavista.com/ (underground search engine)

Issues Regarding Keyword-Based Searches

- Information Overload
- Multiple Vocabularies
- Synonymy
- Polysemy

Issues Regarding Keyword-Based Searches (Cont.)

Information Overload

- The need for effective information retrieval systems becomes increasingly important as computer-based information repositories grow larger and more diverse.
- Information overload is when the users of system are overwhelmed with the amount of current information available, the constant updating of new information, and usually not enough knowledge about the subject and system required to access the information

Issues Regarding Keyword-Based Searches (Cont.)

Multiple Vocabularies

- Users also must deal with multiple vocabularies, which arise from the varying backgrounds and expertise of the users accessing the system. This often leads to poor recall, where the set of data that the user is searching on is to large, or to small, which in turn leads to low precision on the information the user wants.

Issues Regarding Keyword-Based Searches (Cont.)

Synonymy

- Synonymy deals with the equivalence of meaning of words, where different words mean the same thing, or synonyms.
- An example would be searching through the yellow pages for a particular type of business.

Issues Regarding Keyword-Based Searches (Cont.)

Polysemy

- Polysemy deals with the polar opposite problem, where single words have multiple meanings depending on the context in which they are used.

- An example would be searching for the word "pen", which depending on the context used could mean a writing instrument or jail

Alternative To Keyword-Based Search Engines

•Concept-based information retrieval looks at the increasing problem with keyword-based searches with an intuitive approach, which first sorts the data by their relationships, and then searches the data for specific information.

•In sorting the data first by how they are interrelated, we can get a good recall on the set of data that we want, and in searching on this set of related data for specific information, we can get high precision.

• The research into concept-based IR (information retrieval) as a viable alternative is still in its infancy though

Concept-Based Searches: Two Distinct Fields

- Concept-based IR is based upon two distinct fields: Taxonomy and Ontology
- Taxonomy: Division into ordered groups of categories. The science, laws, or principles of classification.
- Ontology: The branch of metaphysics that deals with the nature of being. The relationships between objects in the real world.

Concept-Based Searches: Two Distinct Fields (Cont.)

- Taxonomy deals with the classification of data, where the different objects belong under certain categories depending on their characteristics.
- An example would be the term *laptop*, which would fall under the category of computers→hardware→portable devices.

Concept-Based Searches: Two Distinct Fields (Cont.)

- Ontology deals with the relationship between objects in the real world, such as aggregation ("part-of" relationship such as "an engine is part of an automobile") and inheritance ("is-a" relationship such as "a human is a mammal").
- Ontology answers the question, "What kinds of objects exist in one or another domain of the real world and how are they interrelated?".

Concept-Based Searches: Two Distinct Fields (Cont.)

- The two basic concepts behind ontology are types and roles, where types are instances of an object which always exhibit certain features, while roles are instances of an object which change depending on circumstance.
- An example of a type would be a plant, which exhibits certain features such that it will always be a plant its entire lifetime.
- A role would be a student at a university, where even after they graduate, after they cease to be a student, they are still an individual.

Concept-Based Searches: Two Distinct Fields (Cont.)

- It is important to separate these two fields due to the fact that objects that belong to a certain concept can be classified in very different ways depending on the viewpoint, where the object can be looked at in different ways depending on the user
- The viewpoint corresponds to a role of the object, so there is a one to one mapping between the user and the function of the object.

Concept-Based Searches: Two Distinct Fields (Cont.)

- The way to sort the data in a way which has the advantages of both an ontology and taxonomy is to create a concept map of the current information.
- A concept map is a visual knowledge representation technique, which is used to express relationships between ideas.
- Examples of where concept maps are utilized include brainstorming, planning, documentation, presentation and software blueprints* .

* Whiteboard example

Concept-Based Searches

- A practical example of the concept mapping technique (used for the purpose of relationships between objects) would be the semantic network, where the nodes in the directed graph are the ideas and the links are the relationships between them

- Building such a conceptual network of ideas and objects would give the user a good recall on the information requested (based on idea clusters), and from the information given back to the user, they would choose which of the branches are most relevant to them. Then the user would search the relevant branch of concepts, where the information in the branches is stored in a categorical fashion

Concept-Based Searches (Cont.)

- Currently, there is an effort to sort data on the web (defined and linked) in a way that it can be used by machines - not just for display purposes, but for using it in various applications.
- The name of the effort is “The Semantic Web” (http://www.semanticweb.org/)
- They are employing both the fields of ontology and taxonomy to solve the problems

Concept-Based Searches (Cont.)

- Another example of a concept mapping technique (used for the purpose of relationships between objects) in a search engine would be the Information Mapping Project at Stanford.
- http://www-csli.stanford.edu/semlab/infomap.html (Homepage)
- http://infomap.stanford.edu/webdemo (search engine)

Conclusion

- The volume of information will only increase, so there must be methods established to harvest the data into a coherent form, where a user can find the relevant and interesting information they are looking for with accurate results.
- Combining the two fields of ontology and taxonomy is a good approach to the concept-based model of information retrieval.

Bibliography

1) Taveter, Kuldar. Intelligent Information Retrieval Based on Interconnected Concepts and Classes of Retrieval Domains. http://www.ercim.org/publication/ws-proceedings/DELOS8/taveter.pdf.

2) Rob Kremer, Brian Gaines. Embedded Interactive Concept Maps in Web Documents. Proceedings of WebNet'96: World Conference of the Web Society. San Francisco, CA., 1996. http://www.cpsc.ucalgary.ca/~kremer/webnet96/webnet_kremer.html.

3) Lager, Mark. Spinning a Web Search. California Lutheran University, 1996. http://www.library.ucsb.edu/untangle/lager.html.

4) Hsinchun Chen, Bruce R. Schatz. Semantic Retrieval for the NCSA Mosaic. http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/chen/chenschatz.html.

Bibliography (Cont.)

5) Woods, W. A. Finding Information on the Web. 1995. http://www.ai.mit.edu/projects/iiip/conferences/www95/woods.html.

6) Henninger, Scott. Interface Issues and Interaction Strategies for Information Retrieval Systems. 1996. http://www.acm.org/sigchi/chi96/proceedings/tutorial/Henninger/njb_txt.htm

7) Berners-Lee, Tim. The Semantic Web. 2001. http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html

8) T.B. Rajashekar. Internet and Information Resource Discovery. 2001. http://144.16.72.189/is213/topic-7.htm

End of Presentation

Thank You