CS276A
Text Information Retrieval, Mining, and Exploitation

Project Information and Resources
Project Ideas
3 Oct 2002

Project: The idea
- An opportunity to do a reasonable sized project on a topic of interest related to the course
- Open ended, but...
  - Have a clear, focused objective
  - Set reasonable goals
  - Don’t reinvent the wheel
  - Look for ways to evaluate and validate your work
  - A good project write-up is essential

Project: General Info
- 40% of grade
- Work in teams of two
- Stay after class to look for a partner if needed
- Try asking on news.su.class.cs276a
- TA can help you find a partner
- If you are not registered for the course, you may not have access to some of the data
  - subscribe to cs276a-guests@lists if you are not registered officially
  - send a msg with “subscribe cs276a-guests” in the body to majoradomo@lists

Project: General Info cont.
- Suggested profile information to put on newsgroup when looking for a partner:
  - What times of the day are you free?
  - What kind of project interests you?
    - theoretical
    - efficiency
    - visualization
    - etc...
  - What programming language do you prefer?
  - Do you prefer working remotely, or on campus?

Timeline
- Oct 3, Oct 10: discuss project ideas and resources in class
- Oct 17 (one week later): Hand in project abstract (1 page)
  - We will email feedback by Oct 24
- Nov 12: Submit summary of progress, and plans for completing the project (2–3 pages)
- Dec 3: Submit project
- Dec 5: Present project (very brief “executive summary”) in class

Tools & Resources
Installed/Supported Tools

Linux & Solaris

- Java (J2SE 1.4.0)
- Lucene
- Berkeley DB
- Google API
- WebBase API
- WordNet

- software is in cs276a/software/
- data is in cs276a/data1/

Put the following in your .cshrc:

```
source /afs/ir/class/cs276a/software/setpaths.sh
```

Please take a look at:

- http://www.stanford.edu/class/cs276a/projects/docs/

Running example, making use of many of these tools, is in cs276a/software/examples/

- package cs276a.examples.titleindexer contains a running example we'll be walking through

Other Tools

- Porter’s Stemmer
- Xerces XML parser
- Many free libraries available at http://www.sourceforge.net/

Java

- Rich API of built-in functionality
  - Networking (e.g., HTTP)
  - HTML parser
  - XML support
  - Visualization
  - ...

Lucene

- http://jakarta.apache.org/lucene/docs/
- Library for text indexing
- Free, easy to use, and extensible
- API for both constructing and querying text indexes
  - IndexWriter
  - IndexSearcher
  - Analyzer
- Detailed example later on

BerkeleyDB

- http://www.sleepycat.com/
- External memory databases
  - BTree
  - Hash table
- Might be useful for certain tasks (like link analysis)
- Very widely used
Google API

- http://www.google.com/apis/
- Programmatic interface to Google
- Requires you to register to get a client key
- Limit of 1,000 queries / day
- Supports 3 kinds of queries
  - search - search query (more details next)
  - cache - retrieved cached copy of a url
  - spelling - spelling suggestion

Google API cont.

- Note that "search" query also allows
  - link-queries
    - backlinks of standard: "link:www.stanford.edu"
  - related pages queries
    - similar to standard: "related:www.stanford.edu"
  - intitle:, daterange:, etc...
    - see http://www.google.com/advanced_search

Stanford WebBase

- Recent crawl of Stanford.EDU is in cs276a/data1/stanford.edu/
  - Need to use an HTML parser
- Pages from the Arts, Business, and Computers ODP categories are in cs276a/data1/dmoz/
  - Already converted to plaintext (no HTML markup)
- Use Feeder classes to access
  - Pull-style iterator
  - next0

WordNet

- http://www.cs.cmu.edu/~wn/
- Run 'wnb' to play with the system locally, or use the online Web version
- Library provides programmatic interface to hand constructed semantic relationships for words
  - synonyms
  - hypernyms
  - meronyms
  - etc...
- See glossary of relationships at http://www.cs.cmu.edu/~wn/glossary.html

Datasets

- Find your own data, or use something from cs276a/data1/
- WebBase data, as described previously
- Assorted corpora in data1/linguistic-data/
  - TREC
  - Reuters
  - Excite query logs
  - Bilingual datasets
  - much more...

Machines

- free (solaris)
- elaine (solaris)
- firebird (linux)
- raptor (linux)
- If you switch architectures, and are storing binary files on disk, don’t forget about endianness
**Kerberos trivia**

- “man keeptoken” to find out how to keep your processes authenticated for 1 week
- might be necessary if you are doing complex text analysis and want your process to run longer than a day
- sidenote: processing large web crawls takes up to a few days, even with several 1.4GHz Athlons operating in parallel

---

**Project Ideas**

---

**Parametric search**

- Each document has, in addition to text, some “meta-data” e.g.,
  - Language = French
  - Format = pdf
  - Subject = Physics etc.
  - Date = Feb 2000
- A parametric search interface allows the user to combine a full-text query with selections on these parameters e.g.,
  - language, date range, etc.

---

**Secure search**

- Set up a document collection in which each document can be viewed by a subset of users.
- Simulate various users issuing searches, such that only docs they can see appear on the results.
- Document the performance hit in your solution
  - index space
  - retrieval time
“Natural language” search / UI

- Present an interface that invites users to type in queries in natural language
- Find a means of parsing such questions into full-text queries for the engine
- Measure what fraction of users actually make use of the feature
  - Bribe/beg/cajole your friends into participating
  - Suggest information discovery tasks for them
  - Understand some aspect of interface design and its influence on how people search

Link analysis

- Measure various properties of links on the Stanford web
  - What fraction of links are navigational rather than annotative
    - (how do you tell automatically?)
  - What fraction go outside (to other universities?)
  - What is the distribution of links in Stanford and how does this compare to the web?
  - Are there isolated islands in the Stanford web?

Visual Search Interfaces

- Pick a visual metaphor for displaying search results
  - 2-dimensional space
  - 3-dimensional space
  - Many other possibilities
- Design visualization for formulating and refining queries

Visual Search Interfaces

- Are visual search interfaces more effective?
- On what measure?
  - Time needed to find answer
  - Time needed to specify query
  - User satisfaction
  - Precision/recall
Cross-Language Information Retrieval

- Given: a user is looking for information in a language that is not his/her native language.
- Example: Spanish speaking doctor searching for information in English medical journals.
- Simpler: The user can read the non-native language.
- Harder: no knowledge of non-native language.

Cross-Language Information Retrieval

- Two simple approaches:
  - Use bilingual dictionary to translate query
  - Use simplistic transformation to normalize orthographic differences (coronary/coronario)
- Cross-language performance is expected to be worse. By how much?
- Query refinement/modification more important in cross-language information retrieval. Implications for UI design?

Latent Semantic Indexing (LSI)

- LSI represents queries and documents in a "latent semantic space", a transformation of term/word space
- For sparse queries/short documents, LSI representation captures topical/semantic similarity better.
- Based on SVD analysis of term by document matrix.

Latent Semantic Indexing

- Efficiencies of inverted index (for searching and index compression) not available. How can LSI be implemented efficiently?
- Impact on retrieval performance (higher recall, lower precision)
- Latent Semantic Indexing applied to a parallel corpus solves cross-language IR problem. (but need parallel corpus!)

Meta Search Engine

- Send user query to several retrieval systems and present combined results to user.
- Two problems:
  - Translate query to query syntax of each engine
  - Combine results into coherent list
- What is the response time/result quality trade-off? (fast methods may give bad results)
- How to deal with time-out issues?

Meta Search Engine

- Combined web search:
  - Google, Altavista, Overture
- Medical Information
  - Google, Pubmed
- University search
  - Stanford, MIT, CMU
- Research papers
  - Universities, citeseer, e-print archive
- Also: look at metasearch engines such as dogpile
IR for Biological Data
- Biological data offer a wealth of information retrieval challenges
- Combine textual with sequence similarity
  - Requires blast or other sequence homology algorithm
- Term normalization is a big problem (greek letters, roman numerals, name variants, eg, E. coli O157:H7)

One place to start: www.netaffx.com
- Microarray data
- Sequence data
- Textual data, describing genes/proteins
- Links to national center of bioinformatics
- What is the best way to combine textual and non-textual data?
- UI design for mixed queries/results
- Pros/Cons of querying on text only, sequence only, text/sequence combined.

Peer-to-Peer Search
- Build information retrieval system with distributed collections and query engines.
- Advantages: robust (eg, against law enforcement shutdown), fewer update problems, natural for distributed information creation
- Challenges
  - Which nodes to query?
  - Combination of results from different nodes
  - Spam / trust

Personalized Information Retrieval
- Most IR systems give the same answer to every user.
- Relevance is often user dependant:
  - Location
  - Different degrees of prior knowledge
  - Query context (buy a car, rent a car, car enthusiast?)
- Questions
  - How can personalization information be represented
  - Privacy concerns
  - Expected utility
  - Cost/benefit tradeoff

N-gram Retrieval
- Index on n-grams instead of words
- Robust for very noisy collections (lots of typos, low-quality OCR output)
- Another possible approach to cross-language information retrieval
- Questions
  - Compare to word-based indexing
  - Effect on precision/recall
  - Effect on index size/response time