Issues and Solutions in Pan-China Search

Tom Emerson (湯姆·愛摩森)
Sinostringologist

18th International Unicode Conference, Hong Kong

Software Internationalization Services & Technology

Overview

- Terminology
- Features of Chinese Information Retrieval
- Indexing and Catalog Format
- Query Expansion
- Result Display
- Ongoing Work
- Conclusion

BASIS TECHNOLOGY

Terminology

- Recall
  - The number of documents retrieved out of the total number of documents.
- Precision
  - The number of retrieved documents that are actual relevant to the query.
- DL
  - Digital Library

BASIS TECHNOLOGY

Terminology

- IR and CLIR
  - Information Retrieval and Cross-Language Information Retrieval
- TREC
  - Text REtrieval Conference

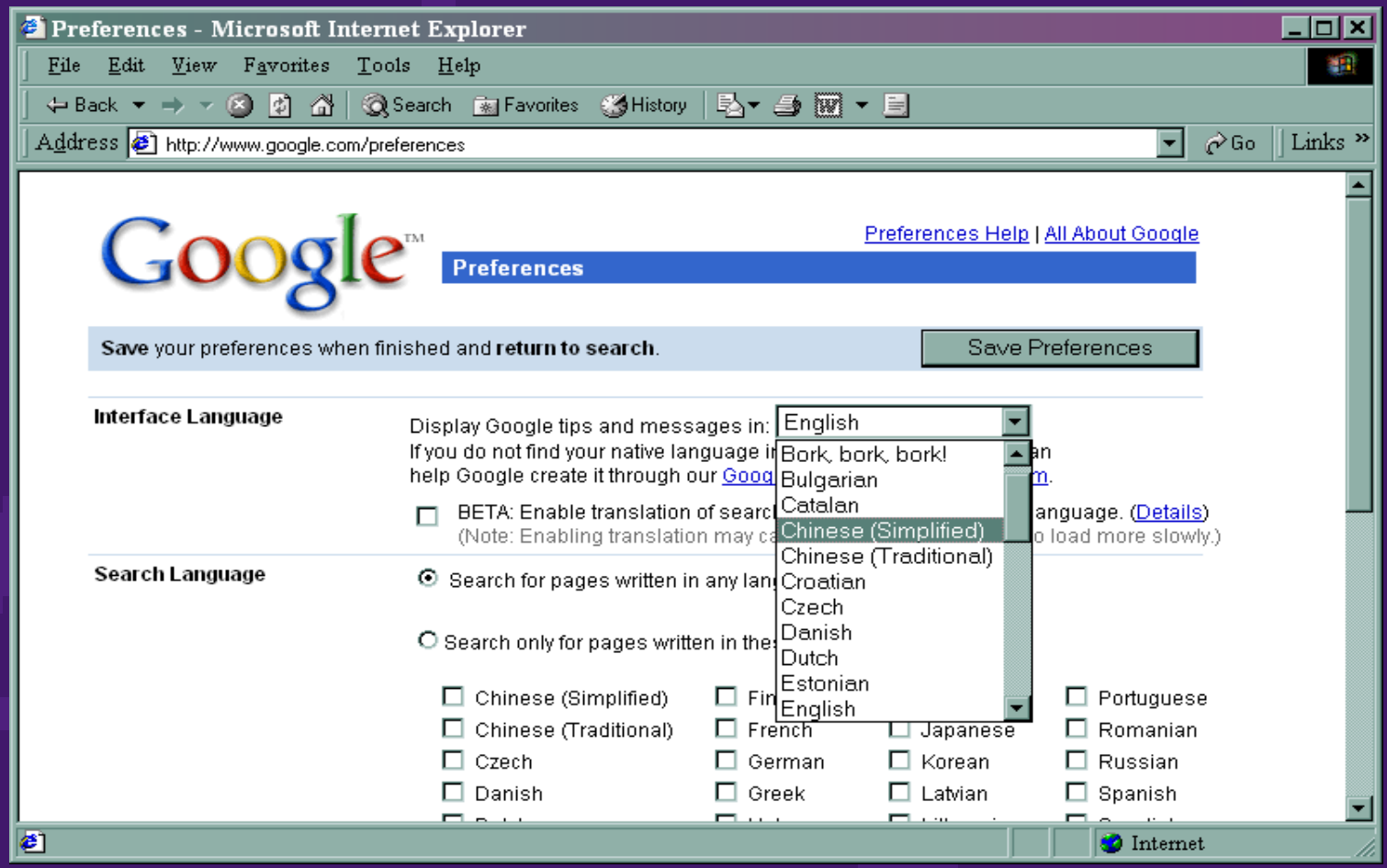
BASIS TECHNOLOGY

Chinese Search

- Domains:
  - The Web
    - Both the Internet and private intranets
  - Corporate document collections
  - Database systems
  - Digital Libraries
- Often limited to a single locale: you can search “Simplified” Chinese documents or “Traditional” Chinese documents.

BASIS TECHNOLOGY

Chinese Search



BASIS TECHNOLOGY

Chinese Search

- Simple Form Characters
  - China and Singapore
    - GB 2312-80, CP936/GBK, GB 18030-2000
- Full Form Characters
  - Hong Kong
    - Big Five, GCCS, HKSCS, CP950, vendor encodings
  - Taiwan and Macao
    - Big Five, Big 5+, ETen, CP950, vendor encodings

BASIS TECHNOLOGY

Chinese Search

- Encoding differences are important, but often overlooked.
- TREC-4 introduced a Chinese tract using Xinhua news articles
- TREC-5 also had a Chinese tract, this time using Hong Kong Government documents
- !!!

BASIS TECHNOLOGY

Chinese Search

- The Browser interface constrains the language/script of the query.
- Form submissions are encoded with the encoding of the containing page.
  - In the case of Google, TC locales use Big 5, SC locales use EUC-CN (aka GB).

BASIS TECHNOLOGY

Chinese Search

- Vocabulary Differences
  - Geographical and political separation lead to differences in vocabulary.
    - 计算机 vs. 電脑
  - Different Chinese locales transliterate foreign names differently.
    - SC 肯尼迪 ken^{3}ni^{2}di^{2}
    - TC 甘迺迪 gan^{1}nai^{3}di^{2}

BASIS TECHNOLOGY

Chinese Search

- Orthographic Variation
  - China defines a standard set of Traditional Characters (STC), some of which are different than those defined on Taiwan (TTC):
    - SC 线 → STC 綫 TTC 線 (xian^4)
    - SC 骂 → STC 罵 TTC 罵 (ma^4)
  - Simplified Characters Used in "Traditional" Locales
    - 台 used instead of 臺 (tai^2)

BASIS TECHNOLOGY

Chinese Search

- Numeric Variation
  - 2000, 二千, 二〇〇〇, 二零零零
- Date Variation
  - 2001-4-27, 2001-27-4, 2001年4月27日
    二零零一年四月二十七日
- These variants need to be unified for effective searching.

BASIS TECHNOLOGY

Chinese Search

- Optimal Chinese search can be viewed as a cross-language IR task (CLIR).
- Unlike generic CLIR, you can expect that the end user can read and mostly understand the results.
  - Hence machine translation (MT) is not necessary, but could be performed

BASIS TECHNOLOGY

Indexing Methods

- Uni- and bi-grams
  - Fast with good recall and decent precision
- Word-based
  - Good recall and precision, providing the segmentation is accurate. Inaccurate segmentation causes a lot of harm.

BASIS TECHNOLOGY

Indexing Methods

- *n*-grams can result in a lot of noise being added to the index
  - Particles such as 了and 的
- Word based methods require a segmentation algorithm
  - Stopwords can be excluded
- Inverted files work very well for either method

BASIS TECHNOLOGY

Catalog Format

- The plethora of locale-specific encodings means you need canonicalize on a single Universal encoding
  - Unicode
  - ISO 2022-CN-EXT
  - GB 18030:2000
- Guess which one people use?

BASIS TECHNOLOGY

Catalog Format

- Most IR engines are written with Western languages in mind.
  - 8-bit code path
  - Space separated, indexable units
- 8-bit paths speak to the use of UTF-8 within the IR system.
  - Perhaps not the best, but it is easy.
- I would minimally use UTF-16 if starting from scratch.

BASIS TECHNOLOGY

Catalog Format

- How do you handle the character differences between full form and simple form characters?
  - Convert everything to full form (!)
  - Don’t do anything, and rely on query processing to generate appropriate differences

BASIS TECHNOLOGY

Query Expansion

- Convert simple form to full form characters, and vice-versa
- Perform lexemic conversion in the correct direction
  - The locale of the original query dictates the directions we expand the query.
  - Four variables to consider:
    - The source script and destination script(s)
    - The source locale and destination locale(s)

BASIS TECHNOLOGY

Query Expansion

- Difficult when using *n*-gram indexing methods
  - Little or no context, so polygraphic hanzi cause problems
    - 干 → 乾, 幹, 干, or 榦
    - You have two choices: pick one, or generate all possibilities

BASIS TECHNOLOGY

Query Expansion

- Accurate hanzi (simple form ↔ full form) conversion also requires a "word" based method. We don't change the word, but use tables built containing the correct conversions for each word.
  - 计算机 ↔ 計算機
- Lexemic conversion cannot be done on *n*-grams, since it is inherently "word" based.
  - 计算机 ↔ 電腦

BASIS TECHNOLOGY

Query Expansion

- Lexemic mappings built from our internal data.
  - Also investigating the LIVAC project at CUHK.
- Semantic expansion through the HOWNET project.

BASIS TECHNOLOGY

When Do You Expand?

- Two options: do it when the index is built, or when the query is processed.
- In CLIR research the processing is done on the query.
- I don’t know which is better, yet. In the end it shouldn’t make a difference.

BASIS TECHNOLOGY

Result Display

- Right now, results are transcoded to the user's selected interface language.
- Some search engines just display character salad, others indicate that you cannot view the results.
- This is definitely the case with Chinese-language results.

BASIS TECHNOLOGY

Result Display

- Transcode the results from Unicode into the local encoding?
  - Local encodings have widely varying repertoires
  - But at least the fonts will (hopefully) be good
- Return Unicode?
  - No data loss due to encoding conversion
  - Fonts are the problem: either they are ugly, or you don't have them installed.

BASIS TECHNOLOGY

Result Display

- This is a problem that is actively being researched within the Asian DL community.
  - Tetsuo Sakaguchi's research using server side conversion and plug-ins, first published at DL '96.
- Other methods:
  - Render the page on the server and send down an image.
  - Intermix text and small graphics for the missing glyphs.

BASIS TECHNOLOGY

Result Display

- Relying on particular viewers or plugins constrains who can use your service.
- Server side solutions require non-trivial hardware and proxy servers to handle the load of a busy site.
- Neither allow for offline viewing.

BASIS TECHNOLOGY

Result Display

- Where does that leave us?
  - Unicode. So put pressure on the OS and browser vendors to provide decent Unicode indexed fonts.

BASIS TECHNOLOGY

Ongoing Work

- Exploring query expansion techniques for Chinese
  - Utilizing our Chinese-to-Chinese conversion technology
- Prototyping in Python
  - Provides an interactive Unicode environment
  - Normalization classes: dates, numbers, etc.
- Attempting to quantify the effects of encoding and lexemic variation in Chinese IR.

BASIS TECHNOLOGY

Ongoing Work

- Eventually will integrate into an existing IR engine, such as SMART or OpenMUSCAT.
- Take part in the TREC Chinese language track.

BASIS TECHNOLOGY

Conclusion

- Searching Chinese robustly is hard
- Query expansion is necessary, unless you can limit yourself to documents originating in a single locale.
- Unicode provides a common character set that can obviate one of the big variables in Chinese IR.

BASIS TECHNOLOGY

Conclusion

- Real systems are being built in industry and academia:
  - Google, Inktomi, Lycos, AOL, MITRE, BBN
  - Verity, FAST, and others
  - University of Massachusetts, University of Maryland
  - City University of New York, Cornell
- Research Systems Abound

BASIS TECHNOLOGY

Conclusion

Slides with notes will be available next week at:

http://www.basistech.com/iuc/

BASIS TECHNOLOGY

Q&A

我謝你們!

BASIS TECHNOLOGY