



Chang, Yun-Ke and Arroyo, Miguel Angel Morales and Spink, Amanda (2007) Multimedia Chinese Web Search Engines: A Survey. In *Proceedings Fourth International Conference on Information Technology, 2007. ITNG '07*, pages 481 - 486, Las Vegas, USA.

© Copyright 2007 IEEE

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Multimedia Chinese Web Search Engines: A Survey

Yun-Ke Chang^a

Miguel Angel Morales Arroyo^b

School of Communication and Information

Nanyang Technological University

Singapore

^aykchang@unt.edu.sg

^bosornoy@yahoo.com

Amanda Spink

School of Information Systems

Queensland University of Technology

Australia

ah.spink@qut.edu.au

Abstract

The objective of this paper is to explore the state of multimedia search functionality on major general and dedicated Web search engines in Chinese language. The authors studied: a) how many Chinese Web search engines presently make use of multimedia searching, and b) the type of multimedia search functionality available. Specifically, the followings were examined: a) multimedia features – features allowing multimedia search; and b) extent of personalization - the extent to which a search engine Web site allows users to control multimedia search. Overall, Chinese Web search engines offer limited multimedia searching functionality. The significance of the study is based on two factors: a) little research has been conducted on Chinese Web search engines, and b) the instrument used in the study and the results obtained by this research could help users, Web designers, and Web search engine developers. By large, general Web search engines support more multimedia features than specialized ones.

Author Keywords

Chinese web search engines, multimedia, functionality.

1. Introduction

Today, high-speed streaming media, video-conferencing, and Web audio are very common for podcasting, Internet radio stations, and online music distribution. The possibilities of digital video broadcasting are also here with us [8]. However, searching for multimedia objects has become intimidating undertaking. Content production in digital format is possible and relatively inexpensive through devices like digital cameras, digital camcorders, and digital voice recorders or mp3. When it comes to find multimedia content in the Web, there is no universal

map that can help users to find content they are looking for [2].

There is a need for a better Web searching tools that allow finding multimedia content; however, multimedia resources present a huge challenge for search engines. Multimedia objects are expanding in number on the Web everyday. More and more Web sites are adding images, audio, and video in form of podcasts, video clips, etc. General studies and statistics on Web searching appear regularly [15]. People looking for multimedia objects usually utilizing general Web search engines that frequently need query terms as input. Web search engines are incrementally offering particular procedures for searching multimedia objects because they have recognized the needs for content in different media, but not ways to find it [2].

There are some initial endeavors to offer multimedia Web searching services like GoFish www.gofish.com, which focus on video content; mSpace (<http://mspace.fm>), which provides tools for organizing an information space to suit a person's interest; podzinger (<http://www.podzinger.com/>), which is an audio and video search engine; and Singingfish (<http://search.singingfish.com/sfw/home.jsp>) for audio and video. Podcast search engines deliver content using two strategies: by searching metadata or by converting the audio to text, which is accessible for indexing and retrieval [7]. Singingfish uses the first strategy, and by 2003, it had indexed and cataloged around 30 million streams and downloadable files, and was adding from 150,000 to 250,000 every week [6]. GoFish uses the same strategy as Singingfish, but focuses on online music stores, obtaining their catalog data, reorganize and standardize [2]. Podzinger uses the second strategy, and allows users to search audio podcast using text queries [7]. Unfortunately, the second strategy is useless when the file do not contain textual content.

2. Related Studies

Studies have examined distinct multimedia features in search engines and examined image, audio and video retrieval [17]. Images had been observed as the most frequent multimedia object searched and had longer sessions than audio and video, and audio queries were the longest [15]. Even with the existence of the multimedia options, many users didn't use them at all when searching for multimedia objects. Non-multimedia queries were shorter than multimedia queries, which enclosed words like "pictures" or "image" [15]. However, it was observed that on the Dogpile.com meta-search engine, Web users accessing diverse conglomerate of content collections had different search patterns. Dogpile users' inspected just the initial pages of results and typed in 2 to 3 words per query, and web, audio, and news queries had shorter queries, but longer sessions. Moreover, more users' seeking images and videos looked for technical staff assistance [14].

In image indexing support, WebSeer and AltaVista photo finder used three main techniques to search multimedia objects: similarity with other images, text in HTML, and category utilization derived from object features. Length, key frame information, bandwidth are helpful to seek audio and video [16]. The future of multimedia search in the Web should take advantage of speech recognition for audio/video classification and word identification for screening terms, document clustering techniques, and image similarity [16]. The majority of multimedia objects are not searchable and the dependency on text continues, but the majority of multimedia objects do not have much helpful text to describe them in contrast to a web page [5]. Wide-ranging analyses on the systems and techniques for Web image retrieval and some prototype reviews have been provided [9].

Technologies for non-English Web searching are not as mature as those developed for searching English content [4]. Multi-language support is not new for major US Web search engines. In 1999 Google offered multimedia search [10]. A major problem with Chinese language is word (character) segmentation. In Chinese, characters cannot be parsed directly like words in English, complicating the extraction of significant elements from the content. The problem of segmentation may be an issue in the use of metadata that describes multimedia objects. In spite of language difficulty, the multimedia searching on Chinese Web search engines provides great opportunities and challenges.

In China there are approximately 130 million Internet users in a population of 1.3 billion. Some 70

percent of them are below the age of 30, single and the only child in the family. On July 19, 2006 it was reported that the number of Web sites in China had rose to a total of 788400, and the average user spent 16.5 hours per week online(MSNBC.com). The Chinese online market is the second in size after the United States market, and China is globally the fastest-rising market. Chinese users are more attracted to entertainment than news. Chinese Internet companies are entertainment oriented. Many major US companies such as Google and Yahoo are still emerging in the Chinese market [1].

3. Research Questions

The research goal was to explore the state of multimedia search functionality on major general and dedicated Chinese Web search engines. The research questions are the following: a) how many Chinese Web search engines presently make use of multimedia searching, and b) the type of multimedia search functionality available.

4. Research Design

4.1 Data collection

We identified 56 Chinese Web search engines. Some 20 general Web search engines provided multimedia support and 36 specialized Web search engines were identified (Figure 1).

The methodology used to evaluate multimedia Web functionality was developed by Tjondronegoro and Spink (2006). Different multimedia data types were used [13, 18]. A qualitative analysis of absence or presence of multimedia search features for each Web search engine was conducted. The multimedia search functionalities examined [3]: text-based search, content-based search, and advanced functionality such as relevance feedback, the use of searching agents, et cetera.

| General Search Engines -- with multimedia searching support: | |
|---|---------------------------------------|
| 1 163 (Image) | 26 oovv |
| 2 Baidu (Image, map, movie, MP3) | 27 onphoto news picture category list |
| 3 Chinasite (Image, music) | 28 GIGDIG |
| 4 google (Image, video) | 29 EBOOK |
| 5 iask (Image, map, mp3, video) | 30 enet cool |
| 6 MSN (beta) (Image) | 31 fkee |
| 7 ODP | 32 gameking cool |
| 8 PKU | 33 games.enet |
| 9 Sina (audio, video) | 34 Go2map |
| 10 sogou (Image, map, music) | 35 gograph |
| 11 Sohu (audio, picture, video) | 36 h3t |
| 12 SoSo (Image, MP3, video) | 37 mapbar |
| 13 Sowang (Image, mobile, movie, MP3, TV program) | 38 mychinamap |
| 14 Timway graphics (Image) | 39 openv |
| 15 Tom (Image, music) | 40 Pogames |
| 16 Yahoo (audio, image, video) | 41 Poptang |
| 17 Zhongsou (Image, MP3, map) | 42 Sogua |
| 18 21CN (movie, music, TV program) | 43 souyo |
| | 44 tvix category list |
| | 45 XinHuanet |
| specialized search engines -- with multimedia searching support: | |
| 19 116 (directory) | 46 ChaoJi |
| 20 120ask | 47 HuaJun OnlineDown |
| 21 17173 | 48 MyDrivers |
| 22 51 ditu category list | 49 PCOnline |
| 23 51Soshu | 50 Skyon |
| 24 9eky | 51 Soft16 |
| 25 oblinkx | |
| specialized search engines -- without multimedia searching support: | |

Figure 1. List of Surveyed Search Engines

| | |
|-----------------|---|
| 163 | www.163.com |
| Baidu | www.baidu.com |
| Chinasite | www.chinasite.com |
| google | www.google.cn |
| iask | www.iask.com |
| MSN | cn.msn.com |
| ODP | http://dmoz.org/World/Chinese_Simplified |
| PKU | http://e.pku.edu.cn/ |
| Sina | www.sina.com |
| sogou | www.sogou.com |
| Sogua | www.sogua.com |
| Sohu | www.sohu.com |
| SoSo | www.soso.com |
| souyo | www.souyo.com |
| Sowang | www.sowang.com |
| Timway graphics | http://graphics.timway.com |
| Tom | http://i.tom.com |
| Yahh | http://cn.yahoo.com |
| Zhongsou | http://www.zhongsou.com |
| 21CN | http://search.21cn.com/index.html |

4.2. Multimedia data types examined

- Graphics (G)
- Phone images (PI)
- Images (I)
- Animation
- Video (V)
- Audio (A)
- Structured audio (S)
- Composite type (C)
- Presentations (P)
- Media elements for authoring (R), e.g. video background, power point templates, photo clips, web elements, tiled background, email stationery, etc.

4.3. Multimedia Search Functionalities Examined

Text-based search - Metadata includes: filename, caption, alt, and other tags (Swain, 1999) etc.

- Cues from the text and HTML source code: Image file names, image captions (description), alternative text (alt), HTML title of the document, hyperlinks, and other text.
 - Cues from image content: color or grayscale, image size, file type, file size, and file date.
- Symbolic queries: precise knowledge – SQL like query.

Content-based search - Interactive browsing: leisure users may not have the specific ideas on the image or video. Instead, they can use image/video icons, key frame.

- Query by Example:
- Image: shape, spatial relation, color, texture.
 - Video: motion of object, spatio-temporal relations.
 - Audio: humming for music, keyword, sampling rate, date of creation, analysis feature vector (consists of duration, pitch, amplitude, brightness, and bandwidth).
- By features and sketch (e.g. weighted color, motion pattern, and shape-orientation).
- Object-appearance (e.g. face recognition)

4.4. Advanced functionality

- Relevance feedback (ask users to tag search results as "relevant", "not relevant", or "neutral" -> the feedback will be used for assisting future queries).
- Search with agent software: high-level search gateway to hide users from complex details (e.g. metasearch)
- Hypermedia enhancements: navigate in content in a non-linear manner (like the hypertext paradigm).
- Knowledge-assisted:
 - Subject-object annotation (or descriptive knowledge)
 - Rule- based (or derivation knowledge)
- Taxonomy of contents:
 - Homo-/hetero- geneity of component media: single media or multimedia
 - Source of contents: contents of value, relationship, and derived semantics.

4.5. Multimedia Features examined:

Keywords (Ky, Tx) - Keywords are still the most intuitive search for most novice and expert users.

- Extent of personalization in this feature: Manual formula (M), All of the words (A), exact phrase (E), any of the words (Y), not related to the words (N).

Size (Md) - Users with lower bandwidth or storage can search for smaller file size. Moreover, for images, users can select the resolutions for phone, or desktop wallpaper.

- Extent of personalization in this feature: Small (S), medium (M), large (L), wallpaper (W)

File Format or Type (Md) - Some users want to rely on quality of some file compression formats. For example, jpeg works well in most of photographic pictures.

- Extent of personalization in this feature: the choice of .jpg, .avi, real, windows, etc.

Taxonomy of contents (Tx, Ss, Ms) - Most users would benefit from searching particular topics, subject, actors, etc. These are usually cues of text in the form of metadata.

- Extent of personalization in this feature: Topics based (T) e.g. sports, news, etc.; Features based (F) e.g. most popular, newly added (featured), filename, etc.

Color (Md) - Some users may prefer to get only full colour photographic photos. This feature is not applicable to audio.

- Extent of personalization in this feature: Black-and-white (B), greyscale (G), full colour (F)

Site- or domain- or Website- type (Md) - Often the content of media can be predicted based on the source, which is the website that contains the media. For example, we may search for images within .com domain to find commercial photos.

- Extent of personalization in this feature: Manual (M), .com (c), .edu (e), .gov (g), .org(o), Open directory (d), about.com (a), business.com (b)

Content filtering (Ob) - Content filtering is important for protecting children and workers from explicit contents. Most of current engines emphasize on filtering adult contents while in the future there should be other filters such as violence, drug abuse, and coarse language.

- Extent of personalization in this feature: Moderate – explicit image (M), Strict – explicit image and text (S), Adult (A), Family filter (F), offensive material (O).

Duration (Md) - Selecting the duration of audio or video allows users with lower bandwidth to limit the search results on the media that they can view quickly with possible streaming. This feature is not applicable to image.

- Extent of personalization in this feature: usually in the forms of minimum or maximum length.

Viewing or Delivery method and some descriptions on the media (Vi, Ss, Md) - Some search engines provide the full-content of the media such as full-sized images and full-length video clips, while most engines simply provide the alternatives for navigation and the content is accessible from the link to the source site.

4.6. Extent of personalization in this feature

- Video: Free-download (D), Pay-per-download (P), Streaming or viewing only (V),
- Image: Full-size available (F)
- Alternatives for navigation: Image/Video thumbnail (T); Media content is described with text high-level semantic captions and low-level captions such as filename, etc. (S); Media content is described with only low-level text such as file name, resolution, album, category and source hyperlink address (L).

The information for this study was collected from August 14 to September 2, 2006. Multimedia features that were analyzed are the following: keywords, size, file format or type, taxonomy of contents, color, site- or domain- or website- type, content filtering, duration, viewing or Delivery method and some descriptions on the media. In Table 1, we use some additional notions: different media is treated as a separate search; and different media is supported with text search with only media type option. Authors compared the extra and more advanced functionalities of video-specialized search engines.

5. Results

We examined 56 Chinese Web search engines. Among them, 20 are general purpose Web search engines and 36 are specialized search engine. Some 19 of the general Web search engines support multimedia, with some components of 7 general search engines not supporting multimedia search and one component of one general search engine did not work. Some 26 specialized Web search engines support multimedia; 7 do not; 2 were inaccessible, and 1 did not work.

Table 1 summarizes the qualitative results of Web search functionality. Each row represents the percentage of Web search engines that support each particular feature. By large, general Web search engines support more multimedia features than specialized ones. In the multimedia data types examined, general search engines do better than specialized ones in audio, images, news images, and video. However, when it comes to looking for software specialized Web search engines do better.

Table 1. Results summary

| Media | K-Word | Size | Format | Taxono | View |
|----------------------------------|--------|-------|-----------------------|--------|-------|
| General Search Engine | | | | | |
| An=10% | M=90% | S=60% | ArtistDirect=10% | F=95% | X=90% |
| A=60% | A=100% | M=65% | AudioLunchbox=5% | T=35% | L=65% |
| I=90% | E=95% | L=65% | Avi=35% | | S=75% |
| Text=95% | Y=65% | W=45% | BMP=45% | | T=65% |
| Icon=5% | N=55% | | CRBT=5% | | V=5% |
| News I=15% | | | FLASH=60% | | D=5% |
| Pl=5% | | | Game=10% | | P=5% |
| V=45% | | | GIF=75% | | |
| C=0% | | | JPG=80% | | |
| P=0% | | | Karaoke=5% | | |
| R=0% | | | Lyric=25% | | |
| | | | Mobile Ring=40% | | |
| | | | mp3=45% | | |
| | | | MPEG=35% | | |
| | | | Music Box=10% | | |
| | | | PNG=45% | | |
| | | | quicktime=5% | | |
| | | | RM=60% | | |
| | | | Sogua Media Player=5% | | |
| | | | swf=10% | | |
| | | | text=45% | | |
| | | | WMA=65% | | |
| | | | Other=50% | | |
| Specialised Search Engine | | | | | |
| An=6% | M=45% | S=35% | ArtistDirect=0% | F=94% | X=36% |
| A=21% | A=97% | M=25% | AudioLunchbox=0% | T=18% | S=40% |
| I=45% | E=94% | L=40% | Avi=21% | | T=48% |
| Text=61% | Y=70% | W=30% | BMP=21% | | F=18% |
| V=30% | N=24% | | CRBT=0% | | D=12% |
| S=3% | | | FLASH=12% | | |
| Sw=18% | | | Game=3% | | |
| | | | GIF=30% | | |
| | | | JPG=27% | | |
| | | | Karaoke=0% | | |
| | | | Lyric=3% | | |
| | | | Mobile Ring=6% | | |
| | | | mp3=3% | | |
| | | | MPEG=18% | | |
| | | | Music Box=0% | | |
| | | | PNG=18% | | |
| | | | quicktime=0% | | |
| | | | RM=30% | | |
| | | | Sogua Media Player=0% | | |
| | | | swf=0% | | |
| | | | text=45% | | |
| | | | WMA=33% | | |
| | | | Other=48% | | |

Animation, icons, phone images are supported by few Web search engines. Physical or logical mixing of multiple basic types, presentations, and media element for authoring are not supported. With keywords, general Web search engines did better in supporting the following: “manual formula” (M), “not related to the words” (N). The results are equivalent for both types of search engines: “all the words” (A) and “any of the words” (Y).

6. Discussion

When it comes to file size, general Chinese Web search engines offer more options than specialized ones. The most common file formats are supported by both types of Web search engines. However, some of them offer aspects that are not common in global Web search engines, such as karaoke resources, mobile rings, and media player.

The possibility to find different images in black-and-white (B), grayscale (G), and full color (F) is not well supported by any type of search engine. In content taxonomy, features based (F) is better supported by both type of search engines than topic based (T). In this last characteristic, general search engines (35 %) do better than specialized one (18 %), see table 2.

Apparently adult content is not a priority for both

general and specialized search engines. The preferred methods to view content are through links and thumbnails. Duration of content was not a feature provided.

The advanced multimedia search features supported by specialized Web search engines are the following: interactive browsing – thumbnail and text summaries; results are sorted by relevance, popularity, and new items added; text alternative uses titles and key-words, media quality was not defined; and the special features are hot ranking, multi-language, complex queries, and Boolean calculations.

Many Chinese general Web search engines search for web pages only. However, they provide multimedia search functions by linking users to special search engines sites that are powered by the same retrieval engine. For example, the popular Chinese Web search engine Baidu has 34 special search engines for different subjects and media types, such as MP3, hand phone entertainment, government sites, maps, etc. The browser opens a new window when users click on a link for specialized search. The interface and display of search results for each specialized search engine can be quite different from each other, too. For example, image.baidu.com gives users further options to browse images in various genres. Users can specify to find picture from news pages which has sub-categories: national, international, social, finance, sports, entertainment, real estate, technology, cars, and Internet. The retrieved images are listed as thumbnails, each one with the URL and the title of the page where the image is embedded, and the name of the web site. On the other hand, searching Yellow Page will result in a simple list of companies and their contact information. Similar to some English search engines, Baidu provides advance search in which users can specify domain, file format, occurrences of key words, currency, etc. In Baidu Web Page search, users can narrow down their search by searching only the current retrieved sites.

7. Conclusion

The retrieval of multimedia content is a great challenge, not only in China, but also at global level. From this study, we obtained insights into the way Chinese Web search engines address the needs of this particular market, such as the resources for entertainment (including game sites), software, drivers, blogs, and others.

Although there is a need for searching tools that can retrieve multimedia content, Chinese Web search engines offer limited multimedia search functionalities. Research in this area has been addressing information retrieval from different angles. The development of

Chinese Web search engines that provide more features for searching multimedia content will generate new trends. The adoption of these trends will depend on how well specialized search engines solve the users' needs. Given the complexity of retrieval problem, text-based search will continue as the most plausible alternative for the near future because of the current practical technology available – metadata and voice recognition.

In the future study of Chinese search engines, the authors found it is necessary to develop a new scheme for classifying Chinese search engines. The conventional definitions of types of search engines may be inadequate when surveying the landscape of multimedia Chinese search engines. In the situation such as Baidu with 34 specialized engines, we can consider www.baidu.com, which search for web pages, as a general search engine, and the rest as individual specialized search engines. We anticipate this type of difficulty will be faced by researchers, for major popular general search engines may eventually dominate the China market, and they evidently have to provide users with capacity in searching for online multimedia resources due to the entertainment nature of the Chinese Internet usage.

8. References

- [1] Barboza, D. (2006). *The Rise of Baidu* (That's Chinese for Google). New York Times. September 17. Retrieved September 20, 2006, from web site: <http://www.nytimes.com/2006/09/17/business/yourmoney/17baidu.html?ex=1316145600&en=66b5bbadeb5c493b&ei=5088&partner=rssnyt&emc=rss>
- [2] Bruno, A. A deeper multimedia search. *Billboard*. N. Y. 117(21), May 21, 2005, 15.
- [3] Chang, S-F., Eleftheriadis, A., and McClintock, R. Next-generation content representation, creation, and searching for new-media applications in education. *Proceedings of the IEEE*, 86, (1998), 884-904.
- [4] Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T., Chen, H. (2004). Internet Searching and Browsing in a Multilingual World: An Experiment on the Chinese Business Intelligence Portal (CBizPort). *Journal of American Society for Information Science and Technology*, 55(9), 818-831.
- [5] Deuel, R. (2004). Multimedia search: ready or not? *IEEE Distributed Systems Online*, 5(7), 1-7.
- [6] Fritz, M. (2003) Singingfish: Advancing the Art of Multimedia Search. *Econtent*, 26(4), Apr 2003, 52-53.
- [7] Gordon, S. (2006) The Search for Podcasts Is On With Podzinger. *Econtent*, 29(2), Mar 2006, pp. 12.
- [8] Johnson, R.B., *Internet multimedia databases. Multimedia Databases and MPEG-7* (Ref. No. 1999/056), IEEE Colloquium on (1999).
- [9] Kherfi, M.L., Ziou, D. and Bernardi, A. (2004). Image retrieval from the World Wide Web: Issues, Techniques and Systems. *ACM Computing Surveys*, 36 (1), 35-67.
- [10] Levander, M. Google Launches Asian Language-Search Tools. *Wall Street Journal* (Eastern edition). Oct 18, 2000. 1. New York, N.Y.
- [11] Mukherjea, S., Hirata, K., and Hara, Y. Towards a multimedia World-Wide Web information retrieval engine. *Computer Networks and ISDN Systems. Sixth International World Wide Web Conference*, 29, (1997), 1181-1191.
- [12] Ozmutlu, S., Spink, A., and Ozmutlu, H. C. Trends in multimedia Web searching: 1997-2001. *Information Processing and Management*. 39(4) (2003), 611-621.
- [13] Pazanda, P., and Srivastava, J. Evaluating object DBMSs for multimedia. *IEEE Multimedia*, 4 (1997), 34-49.
- [14] Spink, A. and Jansen, B.J. Searching multimedia federated content Web collections. *Online Information Review*, 30(6) (2006).
- [15] Spink, A., & Jansen, B. J. (2004). A study of Web search trends. *Webology: An International Electronic Journal*, 1(2) <http://www.webology.ir/2004/v1n2/a4.html>
- [16] Swain, M.J. Searching for multimedia on the World Wide Web. 1999. *IEEE International Conference on Multimedia Computing and Systems*, 1999.
- [17] Tjondronegoro, D., and Spink, A. *Multimedia Web search engine functionality: An exploratory study* (Forthcoming).
- [18] Yoshitaka, A., and Ichikawa, T. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11, (1999), 81-93.