VSM-Based Investigation and Analysis of Navigation Text on Top US University Library Websites: Postprint

Authors: Yin Xiangquan, LI Shuning

Date: 2017-11-08T00:00:00+00:00

Abstract

【Objective】To provide recommendations for navigation construction of domestic university libraries by analyzing characteristics of navigation texts on websites of top-tier US university libraries. 【Method】Following the principle that a world-class university should possess world-class disciplines, faculty, and students, 15 libraries from top-tier US universities were selected. Based on tag clouds and the Vector Space Model (VSM) text mining model, commonalities and specificities of navigation texts were analyzed at the word dimension, with data validation performed using the 2016 State of America' s Libraries Report. 【Results】 Compared with manual investigation, the VSM-based statistical method can yield basic research findings more intuitively and rapidly, which can serve as reference for further in-depth text analysis. 【Limitations】 Only primary navigation, secondary navigation, and homepage title section texts were selected. 【Conclusion】The statistical method based on text data mining models can provide basic research results more intuitively and quickly, offering references for university library website navigation construction.

Full Text

Investigating and Analyzing Website Navigation Texts of Top U.S. University Libraries Based on VSM

Yin Xiangquan, Li Shuning
(Beijing Normal University Library, Beijing 100875, China)

Abstract

Objective: To analyze the characteristics of navigation texts on top U.S. university library websites and provide recommendations for domestic university library navigation construction.

Methods: Following the principle that top-tier universities should have top-tier disciplines, faculty, and students, we selected 15 top U.S. university libraries. Using tag clouds and the Vector Space Model (VSM) text mining model, we analyzed the commonalities and specificities of navigation texts at the word level, and validated our findings against the *2016 State of America' s Libraries Report*.

Results: Compared with manual investigation, the statistical method based on the VSM model can produce basic investigation results more intuitively and rapidly, which can serve as a reference for further in-depth text analysis.

Limitations: Only first-level navigation, second-level navigation, and homepage section titles were selected as the navigation texts for analysis.

Conclusion: Statistical methods based on text data mining models can provide more intuitive and rapid basic investigation results, offering a reference for the construction of university library website navigation.

Keywords: U.S. Academic Libraries; Website Navigation; Vector Space Model; Investigation and Analysis

Classification Number: G250.7

Introduction

In the Internet era, library websites have become the primary window for displaying library content and services, and accessing information through library websites has become a fundamental way for users to utilize library resources. In recent years, as the volume of library information resources and service content has increased—including providing document resources, library service information, library news, and facilitating communication and interaction with readers—the content that library websites must carry has far exceeded original limits. Meanwhile, user needs have become increasingly diversified, requiring them to find what they need on library websites as quickly as possible. However, due to the limited capacity of library websites and the presence of many relatively independent resource management systems and service modules, a considerable amount of content inevitably becomes hidden, preventing library services from being better utilized by readers and leaving readers unaware of what services the library offers. In this context, library website navigation, as the most direct and convenient tool for accessing website content provided to users, becomes extremely significant, and this significance will only grow with information overload. Library website navigation primarily refers to a row of horizontal navigation text located in the webpage header area, either above or below the header banner image, which serves to link various pages of the library website. Additionally, considering that the homepage can also provide a simple and quick entry point to help users rapidly locate needed resources, the titles of various

sections on the library homepage should also be included in the scope of library navigation.

Library websites should improve their navigation functions to purposefully reorganize physically dispersed and disorganized information resources, enabling web users to quickly find the information they need [**?**]. When constructing library websites, major domestic universities have mostly referenced foreign university library websites. Taking Tsinghua University Library as an example, it treated the organization, revelation, and layout of information and resources on the library website as an important issue during its redesign and referenced foreign university library websites [**?**]. However, most major domestic university libraries have focused on describing the implementation after referencing foreign university library websites, without describing or analyzing the navigation investigation methods. Therefore, this paper selects navigation texts from 15 top U.S. university libraries and conducts research using the Vector Space Model (VSM), aiming to provide references for domestic library website navigation construction.

2.1 Data Collection

Following the principle that top-tier universities should have top-tier disciplines, faculty, and students, we selected 10 top U.S. universities [**?**]. Additionally, considering the distinctive social service characteristics of American university libraries, and to examine libraries' social services, we added five more universities to the original 10 [**?**]. The final list of 15 university libraries selected includes: Harvard University Library, Stanford University Library, MIT Library, Yale University Library, Princeton University Library, Columbia University Library, University of Chicago Library, California Institute of Technology Library, University of Pennsylvania Library, UC Berkeley Library, Cornell University Library, UC Davis Library, University of Tennessee Library, North Carolina State University Library, and Northwestern University Library.

A tag cloud is a set of related tags with corresponding weights. The weight affects the font size or other visual effects. The larger the tag font, the more frequently the item appears on the website. Tag clouds are highly suitable for intuitively displaying prominent website content and can be used for commonality analysis among various website navigations. A typical tag cloud contains 30 to 150 tags; when there is too much text, the intuitiveness of the tag cloud decreases.

The text similarity calculation method adopted in this paper is the classic Vector Space Model (VSM). VSM simplifies the processing of text content to vector operations in vector space and expresses semantic similarity through spatial similarity [**?**]. Specifically, the similarity of navigation texts is represented by the cosine value of the angle between two multi-dimensional vectors. The smaller the angle between two vectors, the higher the cosine value, indicating greater

similarity between navigation texts.

In Peter Morville' s User Experience Honeycomb model, findability is one of the main indicators of user experience, primarily manifested through navigation and orientation [**?**]. According to this theory, first-level navigation, second-level navigation, and homepage section titles can represent the main navigation organization method of a website' s content. Therefore, this paper selects first-level navigation, second-level navigation, and homepage section title texts as the analysis objects.

Specifically, assuming there are M navigation texts, we perform feature extraction on each navigation text. Assuming the features are N-dimensional, we can obtain an M×N feature vector matrix F. Mapping this to the VSM model allows us to calculate the feature distances between any two texts among the M texts. This paper first manually collected the first-level navigation, second-level navigation, and homepage section titles of these 15 university libraries, and further performed normalization processing based on the semantics of the navigation text—that is, cleaning and transformation work. For example, "About," "ABOUT," "About us," and "About the library" were unified to "about us" after case unification and text replacement; "Help" and "Get Help" were unified to "get help" ; compound phrases were split, such as dividing "Search & Find" into two terms "search" and "find," and splitting "tools for prospective students|current students|faculty or staff|alumni or friends" into "tools for prospective students," "tools for current students," "tools for faculty or staff," and "tools for alumni or friends." For different levels of analysis objectives, we provide statistical analysis of first-level navigation text, homepage section statistical analysis, and statistical analysis of navigation text at the navigation word dimension.

Corresponding author: Yin Xiangquan, ORCID: 0000-0002-9815-896X, E-mail: yinxq@lib.bnu.edu.cn.

2.2 Text Statistical Analysis Methods

When conducting statistical analysis of texts, we primarily examine the characteristics of individual texts and the similarities between texts. To more clearly parse navigation texts, this paper first uses tag clouds to provide an intuitive overall depiction of navigation texts. Subsequently, statistical methods are employed for feature extraction to characterize each text' s properties, and further, text similarity calculations are performed based on these text representations to mine implicit similarities between texts. Each text' s feature vector can be processed through dimensionality reduction and printed to form summary information for characteristic analysis.

This paper adopts vector cosine distance as the semantic distance between texts. Assuming there are navigation text A and navigation text B, their semantic

distance is shown in Equation (1). dis A B

This paper selects the TF-IDF (Term Frequency–Inverse Document Frequency) method. The main idea of TF-IDF is that if a word or phrase appears frequently in one text but rarely in other texts, it is considered to have good category discrimination capability. VSM-processed texts require operations such as tokenization and stop-word removal. Considering that there are obvious physical spaces between terms in navigation text, this paper treats a set of navigation words as a phrase for processing. First, phrases undergo normalization preprocessing, and each phrase' s TF-IDF value is calculated. Consequently, the representation of each website' s content is converted into a TF-IDF value vector of phrases. Based on this vector, similarity calculations are performed, a similarity threshold is selected, and potential associations among various navigation texts are mined. Additionally, the top 10 features with the highest TF-IDF values for each navigation text are selected and printed as keyword phrases of the navigation text to analyze the characteristics of each website' s navigation.

2.3 Results Analysis

We first used tag clouds to intuitively display the commonalities of first-level navigation text and homepage section titles. Subsequently, to further explore similarities among specific navigation texts, we conducted semantic similarity analysis on each website' s navigation text—including first-level navigation text, second-level navigation text, and homepage section title text—using VSM, printed the feature vectors of each university library' s navigation text, and analyzed them in conjunction with the *2016 State of America*' *s Libraries Report* [**?**].

(1) Intuitive Commonality Analysis of Navigation

After preprocessing navigation terms, we aggregated the first-level navigation text of the 15 top university libraries and imported it into the open-source tag cloud generator TAGUL [**?**] to generate a first-level navigation tag cloud, as shown in Figure 1 [Figure 1: see original paper]. Evidently, "about us," "research support," "services," "get help," "collections," "libraries," "research," "search," and "find" occupy prominent positions on top U.S. university websites, indicating that these navigation terms appear frequently in the first-level navigation of university libraries.

Figure 1. First-level navigation tag cloud of 15 top U.S. university libraries

Similarly, using homepage section title text, we generated a homepage section title tag cloud, as shown in Figure 2 [Figure 2: see original paper]. The modules "news," "search," "find," and "events" are relatively common.

Figure 2. Homepage section titles of 15 top U.S. university libraries

(2) VSM-Based Navigation Commonality Analysis

The VSM-based cosine similarity calculation results for the websites show that similarity between various websites is low (all below 0.30). This indicates that there is no excessive borrowing among university libraries when constructing website navigation, which aligns with actual circumstances and demonstrates the reliability of the text similarity calculation method selected in this paper. Meanwhile, as shown in the intuitive commonality analysis of navigation text, some navigation terms appear in the navigation of multiple university libraries. To further explore commonalities between websites, we selected a similarity threshold of 0.20 and found seven pairs with weak similarity. For intuitive representation, we connected universities with certain similarities using solid lines, as shown in Figure 3 [Figure 3: see original paper]. Harvard University Library shows weak similarity with Stanford University Library, MIT Library, etc. From the printed feature representation vectors, navigation terms such as "libraries," "events," and "staff directory" contribute to the similarity values.

Figure 3. Six university libraries with weak correlation to Harvard University Library and their similarity scores

(3) VSM-Based Navigation Characteristic Analysis

To further examine the characteristics of each university' s navigation text, this paper printed the navigation phrase feature vectors for each university. According to the TF-IDF values of terms from high to low, we selected the top 10 ranked feature phrases for manual analysis, as shown in Table 1 .

Commonalities (considering feature terms existing in more than 1/3 of universities as commonalities): Among the 15 universities, 11 universities feature "libraries" as a key display object, with "libraries" being the first feature term for 10 universities; search-related feature terms are prominent in 7 universities; 6 universities have research guide feature terms ("guides"); 6 universities have event notifications ("events"), with two displayed as "news and events"; 6 universities prominently provide staff directories ("staff directory"); and 5 universities have "about us" feature terms. Through these common navigation terms, it is evident that top university libraries mostly place displaying resources (libraries, staff, about us) and services (search, research guides) in prominent positions.

Table 1. Top 10 navigation text of 15 top university libraries

University	Top 10 Navigation Text
Stanford University	libraries, access for persons with limited mobility, jobs, collecting areas, events, computing(equipment & services), privileges, search tools, chat, course guides

University	Top 10 Navigation Text
Cornell University	faq for instructors, research data management services, equipment, computing, how to submit course reserves, library spaces, help, search tips: catalog, research guides, search
Yale University	libraries, guide to using special collections, get it @yale (borrow direct, interlibrary loan, scan & deliver), find ejournals by title, search worldcat, policies, elischolar, search, search library catalog (orbis), services for persons with disabilities
Columbia University	recommend a title for purchase, deposit your research, technology, borrow direct, interlibrary loan, computing, policies, butler library lockers, room reservation, study spaces
UC Berkeley	libraries, hours and maps, reserve a study room, news and events, renew, staff directory, research help, how to find, online exhibits, about us
University of Chicago	libraries, copyright info, borrowdirect, employment, database finder, chapters, privileges, other local collections, research centers, library surveys
NCSU	libraries, google scholar, interlibrary loan, staff directory, filmfinder, tours, search, give to the library, course reserves, visitor information
Princeton University	libraries, resources, contact the library, tools for alumni, events, interlibrary loan, search tools, visit, guides, tools for graduate students
Caltech	libraries, today' s hours, borrow direct, study spaces and lockers, news and events, recommend a purchase, staff directory, new catalog, about us, research guides
University of Tennessee	publish on demand, site map, archives, how do i··· ?, caltech open access policy faq, software available, about us, ask a librarian, friends of the caltech libraries, copyright support
Northwestern University	libraries, initiatives, archives, contribute your research, events, e-resources, resources for alumni, departments, staff directory, get it services
University of Pennsylvania	ut dissertations, employment, give to the libraries, the library society, staff directory, libraries a-z, renew items, citing sources, music library, research guides
MIT	libraries, resources, penn' s libraries, create a video, staff directory, subjects/collections, tools, tutorials for tools, search, digitalpenn

University	Top 10 Navigation Text
UC Davis	scholarly publishing, galleries, tip faq, events, use policy, more search options, your account, about us, citation software, study spaces libraries, melvyl, request a book/article, subject guides, digital scholarship, engineering, borrowing/circulation, carlson health sciences, about us, exhibits

Note: The table appears to have some formatting issues in the original, with university names and content misaligned in some rows. The translation preserves the content as closely as possible to the original.

Unlike domestic universities, these 15 top U.S. university libraries mostly separate event notifications from news and place event notifications in more prominent positions.

By observing the top 10 dimensional feature vectors of each website's navigation in Table 1, this paper summarizes some of the characteristic services they reflect, as follows:

1. In Stanford University Library's feature vector, the second-ranked feature term (phrase) is "access for persons with limited mobility," demonstrating the website's support services for people with disabilities; the sixth feature term, "computing (equipment & services)," shows its support for computing-related services.
2. Similar to Stanford, Yale University Library also reflects services for people with disabilities ("services for persons with disabilities"). Additionally, Yale's Orbis online library catalog system appears in the feature vector.
3. Harvard University Library and Northwestern University Library each have prominent information for alumni, with Harvard reflected in the eighth feature term and Northwestern in the fourth.
4. UC Davis Library's online library catalog system Melvyl appears in the second position. Additionally, the eighth feature term ("carlson health sciences") reflects its health-related characteristic data.
5. MIT Library's scholarly publishing service appears as the first feature term in the vector.

These feature vectors indicate that each university library has its service focus, meeting the technical needs of different users and fields, such as Stanford University's computing-related services, Yale University's Orbis online library catalog system, Princeton University's new catalog system, UC Davis's Melvyl online library catalog system, UC Davis's health-related data, and MIT's scholarly publishing services. Additionally, each university library has its own characteristics in developing social services, including services for people with disabilities and alumni services.

The above results also corroborate the statements in the *2016 State of America's Libraries Report*: libraries are actively transforming services to meet users' technical needs [**?**].

Furthermore, the report mentions that surveys show students and faculty recognize the value of university libraries in demonstrating research techniques, enhancing student literacy, and managing course resources. University libraries are exploring innovative methods to motivate student success through technology communities and digital scholarship centers [**?**]. The above content can also find corresponding data support in this paper' s data. Research techniques: all university libraries are involved, specifically manifested as help, guides, catalogs, search, and find; course reserves: Cornell University (5th), NCSU (9th), Stanford University (10th); publishing on demand: Caltech (1st), MIT (1st); digital scholarship: UC Davis (5th).

In summary, through feature vectors, we can quickly understand the commonalities and specificities among navigation texts of various university libraries and further provide a data foundation for analyzing libraries' current priorities and development directions. It should be noted that feature vectors represent representative texts of website navigation rather than complete texts. For example, in complete website navigation texts, besides Stanford and Yale, NCSU, UC Berkeley, Columbia University, and Cornell University all mention disability-related services, but only the former reflects this in the Top 10 navigation text.

2.4 Implications for Domestic University Library Website Navigation Construction

The investigation results offer the following insights for domestic university library website navigation construction:

(1) When constructing website navigation, each university library can apply, but is not limited to, the following common navigation terms: for first-level navigation, consider selecting common navigation terms such as "about us," "research support," "services," "get help," "collections," "libraries," "research," "search," and "find" ; for homepage sections, common sections such as "news," "search," "find," and "events" can be selected; throughout the entire navigation text, terms like "libraries," "search," "research guides," "events," "staff directory," and "about us" have certain universality and can be selected.

(2) During navigation construction, excessive borrowing should be avoided. On one hand, library-specific elements should be added to navigation terms. For example, if the library has special catalogs, the special catalog abbreviation can be added to the catalog navigation term. On the other hand, foreign universities can be referenced in distinguishing user categories in navigation, such as people with disabilities, alumni, students,

etc. More importantly, we should adapt to the requirements of the times and focus on expanding characteristic services, including social services and technical services that motivate student innovation.

This paper takes navigation text from 15 top U.S. university library websites as investigation objects, conducts statistics on navigation texts at the navigation word dimension based on the VSM model, intuitively displays the commonalities and specificities of each website' s navigation, and validates the data against the *2016 State of America' s Libraries Report*. Through analysis, we found:

In first-level navigation text, most libraries focus on displaying their resources and services, including "about us," "research support," "services," "get help," "collections," "libraries," "research," "search," and "find"; in homepage sections, "news and events," "search," and "quick links" modules are relatively common. Similarity between overall website navigation texts is low, with significant differences in each website' s characteristic vectors. Under a given similarity threshold, only Harvard University Library shows weak similarity with six university libraries including Stanford University Library and MIT Library. Feature terms such as "libraries," "search," "research guides," "events," "staff directory," and "about us" have certain universality. From each feature vector, relevant social service feature terms can be found, such as disability services, and some feature terms echo the current situation descriptions of university libraries in the *2016 State of America' s Libraries Report*.

The results demonstrate that compared with manual investigation methods, statistical methods based on text data mining models can provide more intuitive and rapid visual analysis results of the commonalities and specificities of each library' s website navigation, serving as a reference for domestic construction of top university library websites.

References

[1] Xiang Liling, Dong Xiaoyan, Qu Baoqiang. Study of the Website Construction of Colleges and Universities Library from the Web Page Design[J]. Journal of the China Society for Scientific and Technical Information, 2004, 23(2): 204-208.

[2] Fan Aihong, Shao Min, Zhao Yang. Discussion and Practice on the Principles of University Library Website Design——A Case Study of the Tsinghua University Library Website Redesign[J]. Journal of Academic Libraries, 2006, 24(3): 38-42.

[3] Fan Aihong, Yao Fei, Jiang Airong. The Features of the New Website of Tsinghua University Library and the Website Survey Analyses[J]. Journal of Academic Libraries, 2011, 29(5): 66-69.

[4] Ye Ying. A Survey on the Top Universities and Their Libraries in the United States[J]. Journal of Academic Libraries, 2002, 20(3): 5-8.

[5] Xie Lijuan, Zheng Chunhou. Social Services of Academic Libraries in USA: Reality and Inspirations[J]. Journal of Library Science in China, 2009, 35(2): 93-97.

[6] Semantic Studios. User Experience Design [EB/OL]. [2016-10-15]. http://semanticstudios.com/user_{{{experience}}}_{{design}}}/.

[7] Vector Space Model [EB/OL]. [2016-10-15]. https://en.wikipedia.org/wiki/Vector_{{{space}}}_{{model}}}

[8] American Library Association. The State of American' s Libraries [R/OL]. [2016-10-16]. http://www.ala.org/news/sites/ala.org.news/files/content/state-of-americas-libraries-2016-final.pdf.

[9] TAGUL [CP/OL]. [2016-10-15]. https://tagul.com/.

Author Contributions

Yin Xiangquan: Proposed research ideas, designed research scheme, conducted experiments, collected, cleaned and analyzed data, drafted paper; Li Shuning: Designed research scheme, revised paper.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors, E-mail: yinxq@lib.bnu.edu.cn.

[1] Yin Xiangquan, Li Shuning. 美国一流大学图书馆网站导航文本.xlsx. Navigation text and VSM results of top U.S. university library websites.

[2] Yin Xiangquan, Li Shuning. WebsiteAnalysis.rar. VSM source code for website navigation.

Note: Figure translations are in progress. See original paper for figures.