Construction and Implementation of a Faceted System for Online Medical Communities in UGC Mining: Postprint

Authors: Zhai Shanshan, Pan Yingzeng, Hu Pan, Xu Xin

Date: 2023-04-01T16:15:53+00:00

Abstract

[Purpose/Significance] Online medical communities constitute one of the primary means for the public to query health information. To address the prevalent issues in facet systems of online medical communities under the current Web 2.0 model—such as low facet dimensionality, shallow hierarchical structure, and unreasonable facet focus terms—this study proposes a network health facet type framework, aiming to improve facet navigation in online medical communities and enhance health information service quality.

[Method/Process] From the perspective of User-Generated Content (UGC), and by integrating user-concerned health information topics with network health information quality evaluation, an 18-category network health information facet type framework is formulated. Using UGC health information from the "All Questions" section of the Youwenbida website as the data source, a prototype of an online medical community facet system is constructed.

[Results/Conclusion] The constructed facet system prototype can, to a certain extent, ameliorate the deficiencies of current facet systems, providing a feasible solution for constructing facet systems for online medical communities under the Web 2.0 model.

Full Text

Construction and Implementation of an Online Medical Community Faceted System Based on UGC Mining

Zhai Shanshan[1], **Pan Yingzeng**[1], **Hu Pan**[1], **Xu Xin**[2] [1]School of Information Management, Central China Normal University, Wuhan 430079 [2]Department of Information Science, School of Economics and Management, East China Normal University, Shanghai 200241

Abstract

[**Purpose/Significance**] Online medical communities represent a primary channel for public access to health information. Addressing the prevalent issues in Web 2.0-based online medical communities—namely low facet dimensionality, shallow hierarchical structures, and unreasonable facet focus terms—this paper proposes a network health facet type framework to improve faceted navigation and enhance health information service quality. [**Method/Process**] From a UGC perspective, we developed an 18-category network health information facet type framework by integrating user-concerned health information topics with network health information quality evaluation. Using UGC health information from the "All Questions" section of the YouWenBiDa website as a data source, we constructed a prototype faceted system for online medical communities. [**Result/Conclusion**] The prototype system effectively mitigates existing deficiencies in current faceted systems, offering a viable solution for building faceted systems in Web 2.0 online medical communities.

Keywords: UGC; Health Information; Medical Community; Faceted System

1. Introduction

Health information refers to information related to medical treatment, disease prevention, health care, rehabilitation, wellness, and health education [1]. In recent years, "Internet + Healthcare" has flourished alongside rapid developments in information and network technologies. By the end of 2017, China's internet healthcare user base had reached 253 million, accounting for 32.7% of all internet users nationwide [2]. The internet's convenience, accessibility, low cost, and vast data volume enable the public to conveniently and quickly obtain health information online. Health channels on major portals, healthcare apps, and specialized medical communities have become the most important sources of health information for Chinese users. However, most Web 2.0 communities currently lack established faceted systems, and those with faceted navigation suffer from low dimensionality, shallow hierarchies, and unreasonable focus terms, failing to provide satisfactory user experiences. To facilitate easier, more convenient, and efficient access to health information, there is an urgent need to construct user-centered faceted systems that progressively refine user needs based on health information data.

Drawing from these challenges, we propose a UGC-based faceted system construction approach, focusing on two key tasks: faceted framework construction and facet focus term determination. In the empirical section, we use data posts from the "All Questions" section of YouWenBiDa website as samples, performing UGC synonym identification to build a prototype system.

2. Related Research

2.1 User-Generated Content (UGC) User-Generated Content (UGC) is an information resource creation and organization model that emerged and rapidly developed in the Web 2.0 era [3]. Academic research on UGC has explored multiple dimensions. Zhan Lihua examined how data literacy and user behavior jointly influence UGC domains of interest [4]. Zhao Yuxiang et al. proposed a UGC conceptual analysis framework based on type theory, enriching research methodologies [5]. Jin Yan developed a UGC quality evaluation model based on sentiment analysis to identify low-quality content for public opinion monitoring [6]. Wang Xiwei et al. applied sentiment analysis to mobile library UGC, demonstrating its potential to improve information service quality [7]. Wan Liyong et al. constructed and empirically tested a conceptual model for educational UGC quality satisfaction, identifying completeness, usability, richness, standardization, and effectiveness as key factors [8]. Jin Yan et al. identified abnormal behaviors to form a user-profile-based UGC quality prediction model [9].

Research demonstrates that UGC contains substantial information value. Current academic focus concentrates on UGC sentiment analysis, short-text keyword extraction, and short-text quality assessment. In online medical communities, users serve as both consumers and producers of health information resources. Extracting valuable information from this unstructured UGC represents a core challenge in text processing.

2.2 Health Information Organization and Navigation Research Information organization transforms information from disordered to ordered systematic states [10], representing a critical step in faceted system construction. Information organization quality directly affects faceted navigation effectiveness. Wang Na identified standard construction, management supervision, technical support, and user participation as key elements of ubiquitous network information organization mechanisms [11]. Hou Guanhua et al. found that digital library navigation structures affect elderly users' emotional experience, perceived usability, and task performance, with cognitive load and navigation structure jointly influencing emotional experience and reading performance [12]. Wang Ruojia et al., using log mining methods, proposed a user-centered website design model from query and click behavior perspectives [13]. Chen Guo et al. built a faceted navigation system for the DXY cardiovascular forum by integrating UGC perspectives, conceptual associations, and knowledge bases [14]. Hu Qian et al. proposed a user-oriented industry information resource aggregation model by analyzing how industry users affect information resource aggregation [15]. Zhang Xin et al. developed a faceted classification theoretical model by classifying online health information queries along generic facets and attribute characteristics dimensions [16]. Qiu Minghui synthesized a knowledge system for faceted navigation design in information retrieval systems and proposed systematic design recommendations [17].

In summary, research on information organization and navigation has focused on text content analysis, user behavior studies, influencing factor analysis, or optimization of organization methods. However, health information organization and navigation research remains insufficient, lacking systematic comparison of organization methods and offering limited navigation strategies, resulting in underdeveloped and underutilized faceted systems. This paper proposes a faceted system construction scheme from a UGC perspective and validates its feasibility through a prototype system.

---

3. Faceted System Model Design

3.1 Research Framework Faceted systems, also known as faceted search or faceted query, arrange information according to "facet-subfacet-category" rules, enabling users to narrow, expand, or redirect queries to support interactive and exploratory retrieval behaviors. Building online medical community faceted systems requires addressing two key issues: (1) faceted framework construction, and (2) facet focus term determination. Our strategy involves: (1) extracting basic facet type frameworks through user health information concerns and network health information quality evaluation, and (2) establishing conceptual associations between user network terms and subject terms through UGC synonym identification, combining "CMeSH subject terms + knowledge base + electronic medical records" to determine facet focus terms. The overall research framework is shown in Figure 1 [Figure 1: see original paper].

3.2 Faceted Framework Construction 3.2.1 Facet Type Extraction

Facet types describe facet terminology. Under the Web 2.0 model, users freely ask and answer questions in online medical communities, generating massive short-text health information. Traditional knowledge base-based facet type extraction tends to be overly formal and struggles to associate with unstructured UGC, making it unsuitable for online medical communities. Shang Lili et al. found that health risk, diet, medication, physical activity, and cancer topics receive high user attention in WeChat public accounts [19]. Our investigation of 12 online medical communities revealed that users most directly concern: whether they have a disease, what disease, severity, treatment methods, appropriate medications, treatment risks, disease control, and domain experts. Therefore, identifying user-concerned health information topics is crucial for facet type extraction. Additionally, as communities are primary health information sources, network health information quality directly affects user search experience and health literacy [19].

Our facet types combine user health information concerns with network health information quality evaluation, covering four dimensions: disease, patient, needs, and post information quality, with 18 categories total. "Disease" includes disease names, complications, and symptoms—where disease is the most

fundamental characteristic, though identical diseases may present different symptoms significantly impacting diagnosis and care, and some diseases require differential treatment based on patient group characteristics. “Patient” includes gender, age, medical history, and allergens to clearly and accurately define user conditions. “Needs” include diagnostic methods, treatment approaches, daily care, doctor recommendations, and hospital recommendations—encompassing specific diagnostic/treatment tools and implementing entities. “Information quality” adopts four indicators: reply presence, authority, usefulness, and timeliness, based on domestic network health information quality evaluation research [20-21]. These 18 categories form the basic framework for network health information facet types (see Table 1).

3.2.2 Facet Display Strategy

Users typically query using disease, symptom, medication, and personal information, making these suitable as initial facets. To enhance usability and user-friendliness for other facets, display control is necessary. User behavior patterns exhibit continuity, allowing historical behavior logs to predict future actions [22]. First, when user input terms are overly broad, sub-facets are dynamically added based on actual needs. Second, since facets consist of terms with hierarchical, associative, and equivalent relationships, hierarchically-related facets are arranged by superior-subordinate relationships. Finally, when sub-facet focus terms are few, path depth issues are mitigated by elevating sub-facet focus terms to the current level.

3.3 Facet Focus Term Determination 3.3.1 UGC Synonym Identification

UGC synonym identification establishes conceptual associations between user network terms and subject terms, replacing overly formal subject terms with user-friendly language. UGC synonyms include abbreviations and variant synonyms. Abbreviations include Chinese and English forms (e.g., “AIDS” for “acquired immunodeficiency syndrome”). Variant synonyms represent different expressions of the same concept (e.g., “acquired immunodeficiency syndrome,” “AIDS,” and “acquired immune deficiency syndrome”).

Variant synonym identification primarily relies on existing knowledge bases. Different knowledge bases often use different conceptual expressions for the same term, but their concept synonym sets likely contain repeated terms. The basic approach: when identical subject terms appear in variant concept word collections, they are classified as synonyms. Additionally, since terms have coordinate, broader, and narrower relationships in lexical systems, utilizing these knowledge bases enhances synonym identification probability. For example, the CMeSH medical subject headings mark disease relationships; if two terms share identical broader or narrower terms, they are classified as synonyms.

Abbreviation usage is common in UGC texts. Despite diverse user terminology, patterns exist: most abbreviations extract characters from full terms (e.g., “黑斑

息肉病" from "黏膜黑斑-息肉综合症"), either continuously or non-continuously. The basic approach: search abbreviations in the subject term table; if unhit, split into single characters and search individually. If hit, classify as synonym. If an abbreviation hits multiple subject terms, pair it with the shortest matched term.

3.3.2 Facet Focus Determination

Focus terms serve as entry retrieval points in faceted systems. Since online medical community users are generally untrained, have low medical literacy, and use colloquial expressions, focus term selection must align with network user habits—using popular terms instead of formal language. Based on Section 3.2's facet type framework, focus terms are determined and categorized. The process involves: (1) extracting modular health information attributes from CMeSH, Baidu Baike, and knowledge bases, categorizing them into the facet type framework; (2) extracting modular health information attributes from electronic medical records and categorizing them; (3) merging these to create a facet type-focus term table. The flowchart is shown in Figure 2 [Figure 2: see original paper].

3.3.3 Focus Term Display Strategy

To avoid page overload from excessive focus terms, display sorting control is implemented. If hierarchical relationships exist among focus terms in the facet type-focus term table, display follows the term table order; otherwise, sorting is by frequency. For example, in the disease facet, diseases with more posts (e.g., Alzheimer's, coronary heart disease) are listed first. For new users, display control can be based on usage frequency of facets and focus terms by community majority users.

---

4. Prototype System Implementation: A Case Study of YouWenBiDa Website

YouWenBiDa (https://www.120ask.com/list/) is a health consultation website ranking first among Chinese health websites, with over ten million daily active users. The site organizes content by medical departments, with each department further divided into All Questions, Reward Questions, Solved, Unsolved, and Zero-Answer sections. This navigation reflects typical Web 2.0 health community faceted system problems: low dimensionality, shallow hierarchy, and unreasonable search terms that fail to meet users' progressive, personalized, and targeted information needs. Using the All Questions section as an example, we constructed a health information faceted system prototype.

4.1 Information Collection and Preprocessing We used Java's jsoup technology to collect posts from YouWenBiDa's All Questions section, capturing

post titles, content, publication time, replying doctor information, and each reply. A total of 9,433 posts were collected and stored with three attributes: title, content, and link. The content field included doctor name, title, specialty, reply content, and reply time. Although the data comprised doctor-user health communication, irrelevant posts occasionally appeared, necessitating preprocessing.

We employed a filtering strategy combining Jieba segmentation tools with manual intervention: (1) segmenting post titles using Jieba; (2) analyzing segmentation results to filter keywords related to medical treatment, prevention, disease, health care, rehabilitation, wellness, and health education; (3) deleting posts lacking these keywords. The remaining 9,000 posts formed the prototype system's data source.

4.2 UGC-Based Synonym Extraction and Entry Term Selection Through website investigation, we obtained health data from YouWenBiDa's disease, symptom, medication, treatment, and surgery databases, plus exercise category data from Baidu Baike. After categorizing and summarizing these with electronic medical record data, we formed the final facet type-focus term table (see Table 2).

UGC synonym identification includes inter-term synonym recognition. First, using Baidu Baike, disease databases, and other knowledge bases combined with CMeSH subject terms, we performed variant synonym identification. For example, "epilepsy" in Baidu Baike is also called "羊角风" or "羊癫风"; in disease knowledge bases, it's categorized into "reflexive epilepsy syndrome," "benign epilepsy syndrome," and "epileptic encephalopathy"; in CMeSH, its narrower terms include "febrile convulsions," "focal epilepsy," "generalized epilepsy," "post-traumatic epilepsy," "myoclonic epilepsy," "Landau-Kleffner syndrome," and "neonatal epilepsy." These sources were integrated for variant synonym identification of "epilepsy."

Second, we associated network user expressions with subject terms. For instance, users colloquially say "肚子痛" or "肚子疼" for abdominal pain, while subject tables use "腹痛." UGC synonym identification classifies these as identical concepts.

For abbreviation identification in YouWenBiDa UGC, we segmented preprocessed posts using segmentation tools. Taking "internal medicine" diseases as an example, we divided 9,000+ posts into five long texts (900-1,000 posts each), imported the facet type-focus term table as a dictionary to assist Chinese word segmentation, and treated unrecognized new terms as candidate abbreviations. Splitting these into single characters for dictionary matching established conceptual associations between UGC abbreviations and subject terms. Partial results for internal medicine synonym identification are shown in Table 3 .

4.3 Prototype System Faceted System Construction The prototype's facet types follow Section 3.2's framework, using YouWenBiDa All Questions

post counts as the data source. The system comprises facet design and focus term selection. For facet design: (1) a search box allows term input when needed health information isn't displayed; (2) initial facets include disease, symptom, medication, and advanced options (containing 10 sub-facets: gender, age, allergen, reply status, doctor title, answer count, update time, doctor recommendation, hospital recommendation, and others), with subsequent facets dynamically presented based on user clicks. For focus term selection: preprocessed posts are segmented, UGC synonyms identified, and terms categorized according to the facet type framework. Hierarchically-related terms are arranged by level; overly broad focus terms become new subcategories, either directly under the basic framework or after second-level categories. Table 4 shows a partial faceted system for Alzheimer's disease.

4.4 Prototype System Retrieval Results Display and Analysis To compare prototype and YouWenBiDa faceted system effectiveness, we conducted comparative retrieval result analysis using Alzheimer's disease as a test case. Figures 3 [Figure 3: see original paper] and 4 [Figure 4: see original paper] show retrieval result screenshots for Alzheimer's disease from YouWenBiDa and the prototype system, respectively. Figure 5 [Figure 5: see original paper] shows prototype results for Alzheimer's complication "depression." Figure 6 [Figure 6: see original paper] shows prototype results for Alzheimer's symptom "memory decline." Figure 7 [Figure 7: see original paper] shows combined results for Alzheimer's with symptoms "memory decline" and "incontinence."

YouWenBiDa's faceted approach exhibits: (1) fuzzy logical hierarchy and chaotic classification—department facets show non-department tags like wellness and weight loss; (2) single facet dimension with deep retrieval paths—requiring two sub-facet levels to reach Alzheimer's disease; (3) mismatched focus terms and unreasonable post classification—results include irrelevant posts like "I'm 25 but my chest looks like an old person's" (see Figure 3 [Figure 3: see original paper]).

The prototype system (Figure 4 [Figure 4: see original paper]) provides complication, related symptom, examination, medication, and advanced option facets, with "More" expanding remaining focus terms and "Multi-select" enabling multiple term selection. Focus terms are sorted by frequency.

Figure 5 [Figure 5: see original paper] shows secondary retrieval results for Alzheimer's complication "depression." Since Alzheimer's and depression are distinct diseases, clicking "depression" updates symptom, diagnosis, and medication facets, retrieving posts containing both terms. Figures 6 [Figure 6: see original paper] and 7 [Figure 7: see original paper] demonstrate combined retrieval results that filter out irrelevant posts, improving result relevance and enabling efficient user-customized retrieval.

5. Conclusion

With the rapid development of "Internet + Healthcare," Web 2.0-based online medical communities will become primary channels for health information seeking. Increasing public health awareness has amplified online health information needs and behaviors, yet current faceted systems remain underdeveloped. Addressing this gap, we proposed a faceted system construction scheme for YouWenBiDa's All Questions section, performing UGC synonym identification based on user health concerns and information quality evaluation characteristics. The prototype system effectively resolves issues of low dimensionality, shallow hierarchy, unreasonable focus terms, and low resource coverage, enhancing user experience.

Our faceted system prototype demonstrates generalizability—its analysis and design processes are domain-independent and applicable to other fields. However, limitations remain for future research, including health text topic extraction visualization and expanded prototype application.

References

[1] Zhang Xinyao. Research content and significance of health information needs[J]. Medicine and Society, 2010, 23(1): 51-53. [2] China Internet Development Report (2018)[EB/OL]. [2019-08-06]. http://www.cac.gov.cn/2018-11/06/c_{1123672145}.htm. [3] Xiao Qiang, Zhu Qinghua. Empirical study on factors influencing UGC sharing willingness[J]. Journal of Intelligence, 2012, 31(4): 138-142. [4] Zhan Lihua. Analysis of UGC user behavior causes—dual perspectives of user data literacy and user behavior context[J]. Information Studies: Theory & Application, 2018, 41(4): 28-32, 37. [5] Zhao Yuxiang, Fan Zhe, Zhu Qinghua. UGC concept analysis and research progress[J]. Journal of Library Science in China, 2012, 38(5): 68-81. [6] Jin Yan. UGC quality evaluation model based on sentiment analysis[J]. Library and Information Service, 2017, 61(20): 131-139. [7] Wang Xiwei, Yang Mengqing, Wei Yanan, et al. Evaluation effect of mobile library UGC based on sentiment analysis[J]. Library and Information Service, 2018, 62(18): 16-24. [8] Wan Liyong, Du Jing, Shu Ai. Empirical study on factors influencing educational UGC quality satisfaction—based on extended ACSI model[J]. China Educational Technology, 2019(3): 72-78. [9] Jin Yan, Sun Jiajia. UGC quality prediction model based on user profiling[J]. Information Studies: Theory & Application, 2019, 42(10): 77-83. [10] Lou Cequn, Duan Yaoqing, Zhang Kai. Fundamentals of Information Management (2nd Edition)[M]. Beijing: Science Press, 2009: 131. [11] Wang Na. Research on information organization mechanisms in ubiquitous networks[J]. Journal of Modern Information, 2018, 38(5): 25-31, 36. [12] Hou Guanhua, Dong Hua, Liu Ying, et al. Empirical study on navigation structure and cognitive load effects on elderly users' digital library experience—taking the National Digital Library as an example[J]. Library and Information Service, 2018, 62(13): 45-

53. [13] Wang Ruojia, Li Pei. Research on user health information retrieval behavior based on log mining[J]. Library and Information Service, 2015, 59(11): 111-118. [14] Chen Guo, Xiao Lu, Sun Jianjun. Construction of faceted navigation system for online communities—taking DXY cardiovascular forum as an example[J]. Information Studies: Theory & Application, 2017, 40(10): 112-116. [15] Hu Qian, Li Jing. User-oriented industry information resource aggregation research—taking maternal and infant health user knowledge community as an example[J]. Library and Information Knowledge, 2018(1): 87-95. [16] Zhang Xin, Wang Dan. Research on user online health information search tasks[J]. Information and Documentation Services, 2017(6): 74-83. [17] Qiu Minghui. Research on faceted navigation design for information retrieval systems[J]. Journal of Modern Information, 2018, 38(10): 78-84, 120. [18] Shang Lili, Wang Tao. Research on WeChat health information attention based on user information behavior[J]. Information Science, 2019, 37(8): 132-138. [19] Deng Shengli, Zhao Haiping. Research on constructing evaluation standard framework for network health information quality from user perspective[J]. Library and Information Service, 2017, 61(21): 30-39. [20] Jiang Wen, Xu Xin, Wu Gaofeng. Automated evaluation of online Q&A community information quality with added emotional features[J]. Library and Information Service, 2015, 59(4): 100-105. [21] Qian Minghui, Xu Zhixuan, Lian Yi. Online health consultation platform information quality evaluation and its branding implications[J]. Information and Documentation Services, 2018(3): 57-63. [22] Hu Changping, Lin Xin. Faceted construction based on thesaurus faceted transformation in scientific literature retrieval[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(8): 875-884.

Author Contributions

Zhai Shanshan: Proposed research framework and design; Pan Yingzeng: Data collection, analysis, and initial manuscript drafting; Hu Pan: Prototype system development; Xu Xin: Provided revision suggestions and finalized the manuscript.

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