Knowledge Discovery Strategies and Models for Text Data in Virtual Health Communities: Post-print

Authors: Mou Dongmei, Jū Yuánhóng, Dai Wenhao, Huang Lili

Date: 2023-08-26T00:00:00+00:00

Abstract

[Purpose/Significance] To analyze and propose knowledge discovery strategies for text data in virtual health communities, and to construct a knowledge discovery model for virtual health community text data.

[Method/Process] By summarizing and analyzing the characteristics of virtual health community text data, corresponding knowledge discovery strategies are developed to address the data mining difficulties arising from these characteristics. Under the guidance of the DIKW hierarchy, a knowledge discovery model for virtual health community text data is constructed based on the proposed strategies. Through the application of computer coding, natural language processing techniques, syntactic analysis, formulation of inference rules, and other methods, the data value enhancement process from free text data to adverse drug reaction intelligence is realized.

[Results/Conclusion] Through empirical research, the effectiveness and operability of the proposed knowledge discovery strategies and model are verified, providing a reference for subsequent theoretical and empirical research on knowledge discovery from virtual health community text data.

Full Text

Preamble

Knowledge Discovery Strategy and Model for Virtual Health Community Text Data

Mu Dongmei[1], Ju Yuanhong[1], Dai Wenhao[1], Huang Lili[2]
[1]School of Public Health, Jilin University, Changchun 130000
[2]Modern Educational Technology Center, Changchun University of Chinese Medicine, Changchun 130000

Library and Information Service, Vol. 62, No. 5, March 2018

Abstract

[**Purpose/Significance**] This study analyzes and proposes a knowledge discovery strategy for virtual health community text data and constructs a corresponding knowledge discovery model. [**Method/Process**] By summarizing the characteristics of virtual health community text data, we formulate targeted knowledge discovery strategies to address the mining challenges posed by these features. Guided by the DIKW framework and based on the proposed strategies, we construct a knowledge discovery model for virtual health community text data. Through computer coding, natural language processing, syntactic analysis, and inference rule formulation, we achieve the value transformation from free text data to adverse drug reaction wisdom. [**Result/Conclusion**] Empirical research validates the effectiveness and operability of the proposed knowledge discovery strategy and model, providing a reference for subsequent theoretical and empirical studies on knowledge discovery in virtual health community text data.

Keywords: virtual health community; text data; knowledge discovery; knowledge discovery strategy; knowledge discovery model

1. Health Information Knowledge Discovery Based on Social Media

Virtual health communities, as typical domain-specific social media platforms, have become increasingly important research subjects as more people share and seek health-related information within them. Drawing from the definition of social media mining [**?**], virtual health community mining can be viewed as a process of representing, analyzing, and extracting actionable patterns from virtual health community data. The rapid growth of data in these communities, combined with advances in data mining technologies and the accumulation of biomedical resources, provides both reliable information sources and technical support for discovering useful knowledge from virtual health community data.

Compared to traditional literature databases and research experimental databases, virtual health communities offer several distinctive advantages: (1) they provide massive data resources for mining; (2) their domain data more accurately reflect users' real-world conditions; (3) their data is voluntarily generated by users; and (4) their data exhibits better timeliness with faster update rates. These characteristics—rapid transmission, broad application, and frequent updates—create a rich foundation for data mining and knowledge discovery, though the valuable information remains deeply buried and requires processing to extract implicit knowledge and wisdom.

The DIKW (data-information-knowledge-wisdom) hierarchy presents a layered transformation process from data to information, then to knowledge, and finally to wisdom through progressive refinement. Therefore, based on the DIKW system's data-to-wisdom conversion process, we can abstract a general methodological model to guide domain users in conducting knowledge discovery research on virtual health community text data.

Current research on health information mining from social media primarily focuses on hotspot topic identification and disease trend prediction. X. Ji et al. [**?**] employed a two-step sentiment analysis method to measure user concern levels about diseases through Twitter message sentiment classification. D. Ghosh et al. [**?**] used LDA (latent Dirichlet allocation) topic modeling and spatial analysis to identify health-related topics on Twitter, using obesity as a case study. R. Mehrotra [**?**] proposed two novel methods to improve LDA topic models for microblog content identification, offering significant enhancements without modifying the underlying LDA mechanism. J. Parker et al. [**?**] presented a general framework for detecting public health trends on Twitter, though limited to previously known diseases using Wikipedia and ICD, thus unable to detect emerging diseases. S. Doan et al. [**?**] introduced a new filtering method to identify influenza-like illnesses (ILI) from 5.87 billion tweets, demonstrating that semantic feature filtering is useful for geographically selecting tweets and can be broadly applied to diseases and symptoms. P. Kostkova et al. [**?**] demonstrated social media's early warning capabilities by tracking disease spread during the swine flu pandemic, showing how real-time updates could improve outbreak detection systems, though location identification remains problematic. S.D. Young et al. [**?**] used real-time social media technology to detect and remotely monitor HIV outbreaks, though the study was constrained by factors such as delayed case reporting and inability to assess relationships between recent infections and risk behaviors. D. Barazani et al. [**?**] proposed a system for point-source outbreak surveillance, but it could only monitor known diseases, not predict novel ones.

In summary, domain users require information from online sources, particularly virtual health community data. However, existing research on virtual health community knowledge discovery primarily focuses on analyzing data content to identify research hotspots, frontiers, and trends. Technically, most studies concentrate on information extraction methods and techniques, with few integrating linguistic theory, natural language processing, domain knowledge, and data mining theory to guide the discovery of implicit knowledge in virtual health community text data. Therefore, this study combines natural language processing, syntactic analysis, topic modeling, and ontology mapping to investigate knowledge discovery in virtual health community text data.

---

2. Knowledge Discovery Strategy for Virtual Health Community Text Data

Platforms such as Twitter, Facebook, Weibo, DailyStrength, and MedHelp generate massive amounts of health-related data daily. This user-generated content concerning disease diagnosis, drug development, and adverse drug reactions holds significant research and application value. However, compared to literature and research data, virtual health community text data is highly unstructured, characterized by: colloquial concept descriptions with high degrees of idiomatic expression, missing characters, and singular/plural confusion; entity semantic relationships expressed through context rather than explicit abstraction; platforms for expressing personal feelings where objective events are mixed with emotional expression, making statements more ambiguous; and substantial knowledge remaining implicit and not directly manifested.

To address these challenges, we develop knowledge discovery strategies for entity recognition, semantic relation extraction, and event detection under the guidance of linguistics, information organization, and computer science theories. These strategies collectively form a comprehensive virtual health community data mining and knowledge discovery framework to guide the analysis and resolution of knowledge discovery problems in this domain.

The knowledge discovery strategy for virtual health communities analyzes the causes and manifestations of these data characteristics. For each feature, we propose targeted solutions by integrating information organization theory, ontology mapping theory, computer technology, and natural language processing techniques, thereby guiding knowledge discovery research on virtual health community text data (see Figure 1 [Figure 1: see original paper]).

The textual nature of virtual health community data arises from information communication patterns and storage requirements, manifesting as unstructured free text stored on the web. To address this textual characteristic, we construct a semi-structured text repository to store virtual health community data. The non-standard terminology, user expression variability, and casual communication style result in colloquial concept descriptions, free relation expression, ambiguous event statements, and concealed knowledge. These manifest as non-professional terminology, diverse concept relation types, disorganized content, dispersed data, and fuzzy knowledge expression. We address these issues through syntactic analysis, grammar rule formulation, relational topic modeling, and domain ontologies. Based on these characteristics and solutions, we propose a grammar-rule-based entity semantic relation extraction strategy for the free expression of entity relations and an event detection strategy for ambiguous event descriptions. This research deepens understanding of knowledge discovery theory, promotes deeper application of knowledge discovery, and improves the speed and efficiency of knowledge extraction from virtual health community free text.

---

3. Knowledge Discovery Model for Virtual Health Community Text Data

Building upon targeted knowledge discovery strategies for virtual health community data characteristics, we must address the core issues in text data knowledge discovery research and clarify the workflow. This requires investigating the knowledge discovery model, validating the scientificity and feasibility of the strategies through examining model components and their relationships.

3.1 The DIKW Hierarchy

American management scientist Russell Ackoff constructed the DIKW hierarchy [**?**], a system concerning data, information, knowledge, and wisdom, where each layer endows the next with certain qualities [**?**]. The DIKW system is shown in Table 1 .

Table 1 : The DIKW System

Level	Category	Purpose	Method/Technology
Data	know-nothing	Data screening, computer coding	
Information	know-what	Database technology, syntactic analysis, entity recognition	
Knowledge	know-how	Ontology mapping, relational topic modeling	
Wisdom	know-why	Data mining technology, expert validation	

According to DIKW theory, knowledge discovery from data to wisdom involves: first, extracting heterogeneous massive data from web pages and virtual health communities; second, performing data screening and cleaning to achieve structuring and modeling; third, conducting information integration, statistical analysis, and synthesis to form knowledge; and finally, mining implicit knowledge to provide personalized knowledge services and decision support wisdom. Thus,

wisdom is obtained through progressive refinement from data to information to knowledge, involving processes of data collection, structuring, natural language processing, semanticization, event detection, and knowledge discovery. Virtual health community text data knowledge discovery precisely follows this DIKW process.

3.2 DIKW-Guided Knowledge Discovery Model for Virtual Health Communities

Under DIKW guidance, virtual health community text data knowledge discovery should proceed through: (1) applying computer and database technologies to obtain virtual health community text data from the web, forming a raw data text repository; (2) using natural language processing for preliminary analysis to create information-rich data; (3) performing syntactic analysis to obtain semantic relations; and (4) refining semantic knowledge to ultimately achieve wisdom. Accordingly, the knowledge discovery model comprises five layers from bottom to top: data layer, natural language processing layer, semantic analysis layer, relation extraction layer, and event detection layer. These layers implement the workflow and technical route of text repository construction, named entity recognition, entity semantic relation extraction, event detection, and knowledge discovery.

Data Layer
The data layer is the foundation, comprising data source selection and text repository construction. It enables acquisition of target data from virtual health communities, converting non-structured web data into semi-structured local data to prepare for mining and knowledge discovery.

Natural Language Processing Layer
This layer implements entity recognition and sentence analysis based on linguistic theory. Entity identification depends on part-of-speech analysis, extracting meaningful nouns as objects. Sentence analysis primarily uses grammatical dependency theory to analyze word positions. This processing yields an initial relational data set.

Semantic Analysis Layer
The complexity of virtual health community text manifests as non-standard word usage and free expression that obscures semantic associations. In this layer, we adopt a top-level ontology-based semantic interconnection pattern to normalize concepts, using domain ontologies to map free text and identify domain concepts, thereby obtaining a domain concept collection.

Relation Extraction Layer
Sentence semantics comprises word meanings and inter-word semantic relations. Syntactic structure and semantic relations are crucial for entity relation extraction. The key to successful extraction lies in developing highly discriminative rules based on sentence semantic characteristics. This layer uses inference rules to identify semantic relations between entities, transforming domain concepts

into concept/relation pairs.

Event Detection Layer
Events, as important information expressions, refer to objective facts where specific people and objects interact at specific times and places [**?**]. For virtual health community free text, event detection extracts event content from texts containing event information. Through semantic mapping across multiple domain ontologies, we discover potential event information from concept/relation pairs, compare it with domain knowledge bases or gold standards to identify implicit domain knowledge, and finally validate reliability and accuracy through domain expert verification, achieving the elevation from knowledge to wisdom (see Figure 2 [Figure 2: see original paper]).

---

4. Empirical Study on Adverse Drug Reaction Knowledge Discovery in Virtual Health Communities

Virtual health communities differ from other social media by containing substantial user-generated health information that reflects authentic user feedback on disease treatment and medication. Among these, drugs and their adverse reactions are primary concerns. Pre-clinical trials have limitations in time and subjects, making it impossible to identify all adverse drug reactions, which can have serious consequences. Clinicians need to understand potential adverse reactions to adjust medications, while pharmaceutical manufacturers must monitor post-market drug safety for improvements. Therefore, timely and accurate identification of adverse drug reactions is an urgent global public health challenge.

Following the DIKW hierarchy and our knowledge discovery model, discovering adverse drug reaction knowledge from virtual health communities transforms text data into adverse drug reaction wisdom. This study divides the process into four stages: adverse drug reaction data text repository construction, entity recognition and relation extraction, event detection, and event confirmation.

4.1 Adverse Drug Reaction Data: Data Acquisition and Text Repository (Data→D)

We empirically validate our model by discovering potential adverse drug reaction knowledge from the virtual health community MedHelp [**?**]. Focusing on posts from the kidney disease section, we designed a MySQL database (mysql-5.6.24-win32) to store posts and processing outputs, considering poster information and content length. After filtering, we obtained 19,929 posts related to adverse drug reactions under the kidney disease & disorder theme.

4.2 Adverse Drug Reaction Information Extraction: Natural Language Processing and Entity/Concept Repository (Information→I)

Named entity recognition in virtual health community text identifies diseases, drugs, symptoms, and side effects [**?**]. Given that medical domain ontologies provide decision support, we use UMLS [**?**] (Unified Medical Language System), CHV [**?**] (Consumer Health Vocabulary), and SIDER [**?**] (Side Effect Resource) to semantically annotate free text, achieving semantic mapping between free text and domain ontologies. We employ the standardized medical knowledge base UMLS and MetaMap tool [**?**] to identify biomedical vocabulary in free text.

4.3 Adverse Drug Reaction Knowledge Acquisition: Semantic Relation Extraction and Relation Set (Knowledge→K)

Following our entity semantic relation extraction model, we first complete named entity recognition by mapping to domain ontologies to identify disease, symptom, and side effect concepts. Then, processing text sentence by sentence, we analyze posts using grammar-based inference rules to automatically extract semantic relations between medical concepts, mining disease-drug-symptom relations to obtain adverse drug reaction information. Figure 3 [Figure 3: see original paper] shows partial results, with the black box marking the drug "muscle relaxant"–adverse reaction "ache" relation pair. Using the SIDER2 database as a gold standard, we filter out known drug-adverse event pairs, presenting only potential new adverse drug reactions not found in SIDER2.

4.4 Adverse Drug Reaction Event Confirmation: Event Detection and Implicit Knowledge (Wisdom→W)

Event detection requires distinguishing actual events from potential events through semantic mapping across domain ontologies. Confirming discovered adverse drug reaction events involves validating results to uncover implicit knowledge. Whether the identified drug-medicalSign relation pairs represent new knowledge requires verification by domain experts and pharmaceutical/animal/clinical experiments. Validated drug-adverse reaction pairs can be added to SIDER as gold standard records, guiding clinical medication and patient self-administration while providing references for future research.

Through progressive data processing, virtual health community posts are refined from noisy free text to standardized adverse drug reaction knowledge and wisdom. Each layer's processed information set provides quality data for higher-level analysis, demonstrating progressive value enhancement that ultimately contributes to clinical medication, patient safety, and reduced adverse drug events.

This study addresses the challenges of virtual health community text data mining by proposing knowledge discovery strategies and constructing a model validated through MedHelp data. The model abstracts the revelation process from social media text to implicit domain knowledge into distinct analytical stages,

guiding information processing at different levels. While applicable beyond virtual health communities, limitations include: (1) method specificity requiring further research; (2) limited domain ontology coverage, addressable through more comprehensive ontology integration; and (3) dependence on UMLS vocabulary, which may not cover all colloquial expressions. As internet and social media technologies evolve, we will refine our model to more effectively excavate and apply domain knowledge from virtual health communities.

References

[?] Zafarani R, Abbasi M, Liu H. *Social Media Mining: An Introduction*. Cambridge: Cambridge University Press, 2014: 16.
[?] Chen Y, Li Z, Nie L, et al. A semi-supervised bayesian network model for microblog topic classification // 24th International Conference on Computational Linguistics. Mumbai: COLING, 2012: 561-576.
[?] Jing Yuecheng. Research on Chinese Social Media Event Detection Based on Rich Linguistic Features. Shanghai: Shanghai Jiao Tong University, 2015.
[?] Zhu Xiaoguang. Research on Microblog Sentiment Analysis Based on Semi-supervised Learning. Jinan: Shandong University of Finance and Economics, 2014.
[?] Ji X, Chun SA, Geller J. Monitoring public health concerns using twitter sentiment classifications // IEEE International Conference on Healthcare Informatics. Philadelphia: IEEE Computer Society, 2013: 335-344.
[?] Ghosh D, Guha R. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, 2013, 40(2): 90-102.
[?] Mehrotra R, Sanner S, Buntine W, et al. Improving LDA topic models for microblogs via tweet pooling and automatic labeling // International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast: ACM, 2013: 889-892.
[?] Parker J, Wei Y, Yates A, et al. A framework for detecting public health trends with Twitter // IEEE/AMC International Conference on Advances in Social Networks Analysis and Mining. Niagara Falls: IEEE, 2013: 556-563.
[?] Doan S, Ohno-Machado L, Collier N. Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses // IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology. Piscataway: IEEE Computer Society, 2012: 62-71.
[?] Kostkova P, Szomszor M, St Louis C. Swine flu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. *ACM Transactions on Management Information Systems*, 2014, 5(2): 1-25.
[?] Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine*, 2014, 63(3): 112-115.
[?] Barazani D, Bjelkmar P. System for surveillance and investigation of disease

outbreaks // 23rd International Conference on World Wide Web Pages. Seoul: Association for Computing Machinery, 2014: 667-668.
[?] Ackoff RL. From data to wisdom. *Journal of Applied Systems Analysis*, 1989, 16(1): 3-9.
[?] Bellinger G, Castro D. Data, information, knowledge, and wisdom. *Anaesthesia & Intensive Care Medicine*, 2004, 15(1): 44-45.
[?] Ma Bin. Research on Key Technologies of Event Relation Recognition. Suzhou: Soochow University, 2014.
[?] MedHelp. http://www.medhelp.org/.
[?] Feng Lizhi. Research on Construction Methods of Large-scale Chinese Clinical Medical Record Corpus for Named Entity Extraction. Beijing: Beijing Jiaotong University, 2015.
[?] Unified Medical Language System. http://cintcm.com/yuyan/content/word/UMLS.ppt.
[?] CHV Wiki. http://consumerhealthvocab.chpc.utah.edu/CHVwiki/.
[?] SIDER. http://sideeffects.embl.de/.
[?] MetaMap. http://metamap.nlm.nih.gov/.

Author Contributions

Mu Dongmei: Conceived the research proposition, designed the study framework, revised critical content, conducted final review and approval, provided funding and supervision.
Ju Yuanhong: Responsible for figure organization and paper revision.
Dai Wenhao: Responsible for figure revision.
Huang Lili: Responsible for research framework organization, literature investigation, data collection, cleaning and analysis, and paper writing.

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