Visual Data Mining in a Q&A Based Social Media Website

Jin Zhang and Yiming Zhao

Abstract Data mining methods and technologies have been applied to different social media environments but seldom applied to narrative information based Q&A sites. This paper aimed to employ visual data mining techniques to address health care consumer terms use behavior in the Yahoo!Answers. Three months of data on the topic of diabetes in the health category of Yahoo!Answers were collected and analyzed. Terms from the collected data set were processed, validated, and classified. Both Multi-dimensional Scaling and Social Network Analysis visualization methods were employed to visualize the relationships of terms from related categories (‘Complication & Related Disease’ and ‘Medication’; ‘Complication & Related Disease’ and ‘Sign & Symptom’). Patterns and knowledge were revealed and discovered from the mapping of terms such as “acarbose might cause a side effect of hives”, “antidepressant may increase the risk of developing diabetes”, “there is a connection between imbalance and birthdefects”, etc. The results of this study can be of benefit to both health consumers and medical professionals.

Keywords Data mining · Social media · Social Q&A · Term analysis · Visual analysis
Introduction

Data mining is a knowledge discovery process that reveals hidden patterns and trends from an investigated data set, illustrates relationships among involved objects, and analyzes data in a holistic way. It is widely used in business, health, information sciences and other disciplines.

Information visualization techniques can project abstract and invisible items or objects in a data set onto a visual and observable space where relationships among the projected objects are displayed and people can explore and interact with them. Information visualization and data mining have a natural connection because they share a common purpose. Information visualization can be employed as an effective means for data mining.

Social media provides an interactive online environment where people can create groups of interests, post and share opinions and ideas, discuss issues and concerns, and exchange relevant information in a variety of formats and ways. Social media not only provides users with an interactive environment but also offers dynamic, rich, and open datasets for researchers to utilize. Social media data has been applied to various domains and it is no surprise that researchers use visual data mining techniques to address a domain problem in a social media environment.

With the development of Web 2.0, people seek information from social media instead of completely relying on experts in the Internet. This phenomenon is so widespread that no one can negate its existence and influence. For instance, Yahoo! Answers is the most popular Internet reference site in America (Alexa 2013) and 16.64% users in Yahoo are using Yahoo! Answers.

Yahoo! Answers is a social Question & Answer (Q&A) site, in which questions are categorized and broadcasted to the community. Any user can answer any question. Visitors to Q&A sites are increasingly seeking answers to a wide variety of questions that are organized under topical categories. Questions and answers from users are organized, archived, and searchable for other users (Rosenbaum and Shachaf 2010). Because of these unique characteristics and natural advantages of social Q&A sites, online Q&A sites are fertile ground for future studies in many aspects (Harper et al. 2008).

Social media has become one of the most popular textual and visual data sources for studying individual behavior and dispersive information. Data mining methods and technologies have been applied to different social media collections such as Flickr, YouTube, Twitter, Facebook, etc. but they have seldom been applied to narrative information based Q&A sites.

There are many technologies and tools to do data mining in social media. Applying some of these social network data mining techniques generates very complex models that are hard to analyze and understand (Ferreira and Alves 2012). Visual data mapping, however, is a simple, efficient, and effective mining technique which can present, understand, and explore complex abstract information by using computing techniques (Robertson et al. 1989). Visualization mapping is often employed to reveal connections and relationships among investigated objects, to do
data analysis, to explore information, to explain information, to predict trends, and to detect patterns (Zhang 2008). This study employed two visualization mapping methods to mine social media data.

The astonishing size of social media communities and great diversity of information exchanged within them make these sites a valuable research setting for understanding the general public’s online information seeking (Kim and Oh 2009). The interactions between users and social media include various user behaviors (Liu et al. 2012). Consumer health informatics is supposed to analyze and understand consumer behaviors and contained knowledge. Social media provides data generated by consumers for researchers to investigate the consumers themselves.

This paper will use data mining technologies, especially the information visualization techniques, to address health consumer terms use behavior. Two visual data mining techniques, Multi-dimensional Scaling (MDS) and Social Network Analysis (SNA), were employed to visually analyze the subject terms and their relationships under the topic of “Complication & Related Disease” of diabetes and its related topics in order to discover underlying patterns. Findings of this study can be used to better understand health consumer term usage behavior and provide a new research method to conduct similar research in consumer health informatics.

**Related Work**

Social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 to allow the creation and exchange of User Generated Content (Kaplan and Haenlein 2010). A classification of social media is offered by Kerpen (2013) which includes communication channels such as blogging and social networking, collaboration channels like wikis, multimedia channels like video sharing, reviews/opinions channels like community Q&A forums, etc. A social media platform, which attracts so many users to seek information and which archives the information, is a rich and huge treasure for researchers to conduct data mine research. Evans et al. (2009) distinguished information seeking in social media into public asking and targeted asking. Social Q&A sites are typical public asking platforms and have become some of the most popular destinations for online information seeking (Shah et al. 2008). Most research studies on social Q&A sites focus on user-generated and algorithmic question categorization, answer classification and quality assessment, studies of user satisfaction, reward structures, motivation for participation, and mechanisms of trust and expertise from social Q&A sites (Gazan 2011).

Social Q&A websites have been studied in the domain of health information. Kim et al. (2008) investigated evaluation criteria people use with regard to online health information in the context of social Q&A forums. Zhang (2010) explored contextual factors of consumer health information searching by analyzing health-related questions that people posted on Yahoo! Answers, a mainstream social Q&A site. A recent paper investigated term usage of consumers’ diabetes based on a log
from the Yahoo! Answers via visualization analysis, and ascertained characteristics and relationships among terms related to diabetes from the consumers’ perspective (Zhang and Zhao 2013).

One of the consequences of widespread use of social media is that a new form of labor that arises: the mining of social media data (Kennedy 2012). The emergence of new systems and services of social media has created a number of novel social and ubiquitous environments for mining information, data, and knowledge (Atzmueller 2012). Textual information has been mined for topic prediction, topic discovery, preference recognition and analysis (Kim et al. 2012), sentiment detection (Zhang et al. 2012), community detection and networks identification (Comar et al. 2012), characterization of real-world events and evaluation of the event relatedness (Lee 2012), and other natural language processing tasks. In addition, sentiment analysis is another interesting dimension of social media data mining. It uses linguistic and textual assessment, such as Natural Language Processing, to analyze word use, word order, and word combinations and thus to classify them into the categories of positive, negative, or neutral (Kennedy 2012). Visual data like picture and video in social media have been used to detect and discover emerging topics (Hashimoto et al. 2012). Liu et al. (2012) constructed a generative probabilistic graphic model to study and explore topics and user preference in large-scale multimedia data from Flickr for photos and YouTube for videos. Wang and Yang (2012) used data from community-contributed media as corpus to construct visual-word based image representation. Data from social media were also compared with mainstream media to determine trending topic predictions in video recommendations (Lobzhanidze et al. 2013).

Data mining via social media has been adopted by many businesses such as tourism and the pizza industry. Majid et al. (2013) obtained user-specific travel preferences from geo-tagged metadata in Flickr to recommend tourist locations relevant to users. He et al. (2013) applied a text mining method to analyze unstructured text content on Facebook and Twitter sites in order to do a competitive analysis in the pizza industry. Other interesting social media data mining research studies include: stock pick decisions based on user-generated stock pick votes; download predictions from YouTube video ratings; and the popularity prediction of a story on Digg (Hill and Ready-Campbell 2011).

The boom of social media also brings many data mining opportunities in politics. Wegrzyn-Wolskaaand Bougueroua (2012) discussed a variety of issues and challenges surrounding the use of SNA and Text Mining methods with political applications. They surveyed the French presidential election trends using Twitter’s discussions.

Due to the popularity of data mining in public social media platforms, an abundance of opportunities have become available in health research (Culotta 2010). Mining social media can provide insight to abnormal patterns of disease and aid in predicting disease outbreaks (Guy et al. 2012). Akay et al. (2013) used self-organizing maps to generate a word list that correlated certain positive and negative word cluster groups with medical drugs and devices. Bian et al. (2012) analyzed the
content of Twitter messages to facilitate early detection of potential adverse events related to drugs.

After data are collected from social media, information visualization techniques like the MDS method can be used to effectively find relations within the data. MDS illustrates the relationships among abstract objects and demonstrate emerging clusters in a data set which is free of any data distributional assumptions (Zhang and Zhao 2013). MDS has been used to: analyze relationships among sports-related keywords in addition to traditional hierarchical clustering methods (Zhang et al. 2009); to investigate obesity-related queries from a public health portal (Health Link) transaction log (Zhang and Wolfram 2009); and to analyze frequently used medical-topic terms in queries submitted to a Web-based consumer health information system (Zhang et al. 2008).

In the MDS space, distances between words indicate their relatedness or strength even if the links between words are invisible. SNA can visualize these links and serve as a supplementary method. It uses the same data as MDS.

In summary, social media data mining has been applied to multiple disciplines, including natural language process, sociology, healthcare, business management, etc.

The investigated social media channels were limited to Facebook, Twitter, YouTube, and Flickr. Research studies using social media data mining for health consumer term usage behavior in a Q&A forum are rarely found in the literature. This study investigates a Q&A forum, analyzes and visualizes the data with SNA and MDS.

Methodology

Research data in this study come from a question and answer (Q&A) site: Yahoo! Answers. Using the search term of diabetes and searching under the category of ‘health’ on the website of answers.yahoo.com, 2604 records were collected from 08/10/2011 to 11/10/2011. Records not related to diabetes were deleted. As a result, 2565 records were reserved. Each record consisted of one question and several corresponding answers about diabetes.

Term extraction software was used to grab and extract keywords in the records. The total number of words in all records was 1,043,158. A stop words list was introduced into the software to filter the meaningless words. The list contained prepositions, conjunctions, auxiliary terms, articles, numerals, interjections and other function words. Finally, 20,000 unique words were collected from 1,043,158 total words.

A term validation process was executed to deal with synonyms like ‘man, men and male’ and different forms of the verbs like ‘absorb, absorbs, absorbed, absorbing’, etc. In the validation process, all the forms of verbs were changed to their regular form, and all the forms of nouns were converted to their original form. Words with no relationship in meaning to diabetes were also deleted.
In a previous study (Zhang and Zhao 2013) a diabetes-oriented schema was produced. The categories (Cause & Pathophysiology, Sign & Symptom, Diagnosis & Test, Organ & Body Part, Complication & Related Disease, Medication, Treatment, Education & Info Resource, Affect, Social & Culture, Lifestyle, and Nutrient) from the schema were identified. These categories were used in this study. The identified categories were related to 2565 records. Each category to be analyzed contained a certain amount of words and their frequency in 2565 records. These words were presented in a high dimensional vector space in which dimensionality was determined by the number of unique words from the identified categories.

MDS works very well in mapping and projecting the relationships of objects in a high dimensional record space onto a two-dimensional or three-dimensional space. People can observe proximity relationships among investigated objects intuitively in a low dimensional MDS display space leading to a better understanding of individual or group differences of the investigated objects (Zhang 2008). MDS was conducted at a term level which enabled users to observe semantic relationships among terms within two identified categories. The software used for MDS analysis was SPSS (Version 20).

In this study, a cross categories mechanism was designed to discover underlying pattern and relationship between categories. Categories of “Complication & Related Disease”, “Medication”, and “Sign & Symptom” from the schema were selected for analysis. In order to explore the inner connection between pairs of categories (“Complication & Related Disease” and “Medication”, “Complication & Related Disease” and “Sign & Symptom”), words from the involved categories were first integrated and combined since the analysis was conducted at the term level.

Input data for MDS analysis are a proximity (similarity or dissimilarity) matrix of investigated objects in a high dimensional space. Its output is a spatial object configuration in a low dimension space where users may perceive and analyze the relationships among the displayed objects (Zhang 2008).

The initial step was to establish a raw term-record matrix where the columns of the matrix were Q&A records extracted from the Q&A forum and the row were terms or words extracted from records. Each category, in fact, corresponded to a term-record matrix. After two related categories were identified, the two corresponding term-record matrices were built, and then these two matrices were integrated and combined into one term-record matrix.

Due to the combination, the matrix had to be revised and adjusted in the following manner. If terms didn’t appear in the same records, they were not grouped together in the MDS space. The columns where the summation was equal to 0 in an individual category were removed from the new combined term-record matrix. After this step, the number of the columns in the matrix decreased. In addition, some low frequency words were removed in the matrix because they made little contribution to the later term analysis. After two removal steps, a new word-word proximity matrix was generated based on a similarity measure.

Equation (1) is the converting method used to generate a word-record matrix where \( a_i \) is the frequency of word \( a \) in the record \( x \) and \( b_i \) is the frequency of word \( b \) in the record \( y \).
After the converting procedure, a final word-word proximity matrix is generated as Eq. (2).

\[
P = \begin{pmatrix}
    b_{11} & \ldots & b_{1k} \\
    \ldots & \ldots & \ldots \\
    b_{k1} & \ldots & b_{kk}
\end{pmatrix}
\]

The proximity matrix \( P \) is a \( k \times k \) symmetric matrix. It serves as input data for the MDS analysis. Here \( k \) is the number of the valid words in the two categories and \( n \) is the number of the valid records in the categories.

In the process of mapping the relationships in high dimensional space onto a low dimensional space, information loss and disparity is inevitable and must be controlled in a tolerable range. Therefore, an evaluation criterion was used to assess the reliability and effectiveness of projecting by MDS. The quality of the projection can be measured by the stress value (S), which is defined in Eq. (3). The smaller a stress S is, the better the relationships among the objects in the low dimensional space reflect the relationships among the objects in the high dimensional space and vice versa (Zhang and Zhao 2013).

\[
S = \left( \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left( f(T_i, T_j) - D(T_i, T_j) \right)^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} D(T_i, T_j)^2} \right)^{1/2}
\]

In Equation (3) \( n \) denotes the number of all terms involved; \( D(T_i, T_j) \) indicates the Euclidean distance between two terms \( T_i \) and \( T_j \) in the low dimensional space; and \( f(T_i, T_j) \) is the similarity between terms \( T_i \) and \( T_j \) in the high dimensional space where subscripts \( i \) and \( j \) are two indexes for \( T_i \) and \( T_j \), respectively. Only if the results were eligible in terms of S stress, were they accepted.

In the final MDS space, related words are close to each other. The MDS space can be rotated and zoomed to find an optimal viewing angle. In this way MDS interacts with users to fit users’ current interests. After several rounds of adjustment, the MDS analysis resulted in a low dimensional space where the projected terms were displayed and observed; relevant terms were clustered and specified; and term relationships within a cluster were discovered and examined.

Several steps were taken to conduct SNA in this study using the data generated by Eq. (2). The first step was to convert raw data into ‘##h’ files by Ucinet 6 software and use these ‘##h’ files as an input dataset in the visual software of NetDraw. The second step was to generate the draft of a word network and use
the thickness of a line to represent the strength of connections between two words which were represented by the value of word-word proximity based on Eq. (2). The third step was to remove the less important relationships and corresponding lines from the draft networks. If the value of word-word proximity was less than 0.3 in Equation (2), its related lines and nodes in draft networks were deleted. The final step relayed out the network using the spring embedding algorithm and deleted isolated nodes. In the final words network, words with strong connections were reserved.

Results and Discussion

The visual data mining on texts harvested from Yahoo!Answers was done at a term level. MDS and SNA visualization data mining techniques were employed to process words from the category of “Complications and Related Disease” with words from “Medication” and “Sign & Symptom” respectively in order to explore the semantic relationship among terms and to discover potential knowledge and patterns behind the relationship among terms. The category of “Complications and Related Disease” contained 91 words, the category of “Medication” contained 74 words, and the category of “Sign & Symptom” contained 95 words, initially.

‘Complication & Related Disease’ and ‘Medication’

In the combined category, the Cosine similarity measure was used to create the term-term proximity matrix. In the MDS analysis, the Minkowski distance measure was used and the Minkowski power was equal to 1, the resultant stress value was 0.01959, and the corresponding RSQ was equal to 0.99880. In Fig. 1, a term followed by a “#” sign indicates it was from the “Medication” category while a term followed by no “#” sign means that it was from the “Complication & Related Disease” category.

In Fig. 1, three big clusters emerged. Cluster 1 included antidepressant medicines (citalopram, Effexor, fluoxetine, and abapent in) and medications for type II diabetes (glucophage, glucovance, and onglyza). Cluster 1 also included complications and related diseases (acanthosis, anaemia, cataracts, hives, infertility, ketosis, and polyphagia). Cluster 2 did not have any salient patterns. Cluster 3 covered a majority of words from the two categories. Importantly, as shown in Fig. 1, the “Complication and Related Disease” category and “Medication” category do not have many internal connections under the topic of diabetes in the Yahoo! Answers site. Many unrelated words mingled and some patterns were obscured in the chaotic word clusters.

Figure 2 is the words network of the categories “Complication & Related Disease” and “Medication”. Significant combinations of words were found, including
Fig. 1 The MDS display of categories of “Complication & Related Disease” and “Medication”

Fig. 2 The words network of categories “Complication & Related Disease” and “Medication”
hives from “Complication & Related Disease” and acarbose from “Medication”; fibromyalgia from “Complication & Related Disease” and citalopram from “Medication”; ED, infertility from “Complication & Related Disease” and Viagra from “Medication”; Crohn from “Complication & Related Disease” and prednisone from “Medication”; hypertension from “Complication & Related Disease” and atenolol from “Medication”; and anorexia, bulimia from “Complication & Related Disease” and laxative, magnesia from “Medication”.

To be specific, the first combination of “hives” and “acarbose” brought to our attention that acarbose, as an anti-diabetic drug used to treat type 2 diabetes mellitus and prediabetes, might cause a side effect of hives. This is one of the main findings we got by mining a social Q&A site under the topic of diabetes.

Inspection of the combination of “fibromyalgia” and “citalopram” led us to notice the relationship between fibromyalgia syndrome (FMS) and depressive syndromes since the citalopram is an antidepressant. This is confirmed in a study claiming that emotional depression is common earlier in life and/or at onset of the FMS (Anderberg et al. 2000). Revisiting the questions and answers in the original records, it was found that when diabetics suffered from FMS, citalopram was mentioned in the answers a few times.

What need further discussion are the two relatively large sub-networks in Fig. 2. One noticeable sub-network was comprised of flu, fungus, aneurism, pneumonia, and arthritis from “Complication & Related Disease” and aspirin, ibuprofen, Tylenol from “Medication”. The other remarkable sub-network contained mental, schizophrenia from “Complication & Related Disease” and antidepressant, antipsychotic, Effexor, fluoxetine, Lexapro, olanzapine, Seroquel, Zoloft from “Medication”. These drugs are usually used to treat depression or psychosis. It is worth mentioning that diuretic has a connection with olanzapine and previous studies showed that olanzapine could cause a side effect of urinary retention which is an indication of a diuretic (Deshauer et al. 2006). Furthermore, olanzapine may precipitate or unmask diabetes in susceptible patients (Koller and Doraisswamy 2002). Other antidepressants may also increase risk of developing diabetes (Rubin et al. 2008).

Some discoveries in this research have verified other user-generated content of social Q&A sites. These include the connection of Viagra, a prescription drug for the treatment of erectile dysfunction (ED); prednisone used to treat Crohn’s disease; and atenolol currently recommended only in special circumstances as complementary medication in hypertension (Wikipedia contributors 2013).

‘Complication & Related Disease’ and ‘Sign & Symptom’

In the combined category, the Cosine similarity measure was used to create the term-term proximity matrix. In the MDS analysis, the Minkowski distance measure was used and the Minkowski power was equal to 1. The resultant stress value was 0.16458 and the corresponding RSQ was equal to 0.90483. In Fig. 3, a term followed by a “#” sign indicates it was from the “Sign & Symptom” category while a
term followed by no “#” sign means that it was from the “Complication & Related Disease” category.

In Fig. 3, four meaningful clusters were identified. In Cluster 1, mental, twitch, collapse, and weak were grouped together. Cluster 2 indicated the signs of polydipsia which are related to diseases of polyphagia and polyuria. In Cluster 3 blindness and retinopathy from “Complication and Related Disease” category were connected to blurry, vision, and floater from “Sign & Symptom” category. With the exception of these three clusters, many words from the two categories gathered chaotically in the Cluster 4 and it was hard to find any distinct patterns.
Visual displays were constructed using SNA words network to reveal more hidden patterns in the relationship of categories “Complication & Related Disease” and “Sign & Symptom”. In Fig. 4, not only the patterns revealed using MDS were found, but also other novel and important connections. Firstly, a connection between imbalance (from the “Sign & Symptom” category) and birth defects (“Complication & Related Disease” category) was found. Secondly, headache from “Sign & Symptom” was related to migraine, aneurism, tumor and cancer from “Complication & Related Disease”. Thirdly, schizophrenia joined into a words group found in Cluster 1 of the MDS visualization (Fig. 3). Moreover, cataracts from “Complication & Related Disease” and sight from “Sign & Symptom” were grouped into a word set which emerged in Cluster 3 of the MDS visualization (Fig. 3). That is because they are all related to eyes. It can also be observed that all three complications (blindness, retinopathy and cataracts) had stronger connections (thicker lines) to floater than other words from the “Sign & Symptom” category. Furthermore, lines among polydipsia, polyphagia and polyuria are the thickest in the whole network, which implied their relatedness is the strongest over all the other words. All words related to eyes from both categories were connected to polydipsia, polyphagia and polyuria. Signs such as hunger, thirst, tiredness, numbness, fatigue, dryness, vomit, slow healing were connected because they are the most typical symptoms of diabetes and tend to be listed together by users in a Q&A site.

Fig. 4 The words network from categories of “Complication & Related Disease” and “Sign & Symptom”
Conclusion

Social media makes up a large percentage of the content available on the Internet (White et al. 2012). Users from different backgrounds, different domains and with different requirements collaborate to construct the knowledge treasure together (Chai et al. 2009). Precisely for this reason, researchers are able to mine massive data. Ease of access to such wide user-generated databases and their openness of the contents in social media sites create unprecedented opportunities for researchers to aggregate consumer-think for better understanding of health consumers and how they use the content they generate.

In this study, the visualization data mining methods of Multi-dimensional Scaling and Social Network Analysis were employed to conduct consumer health informatics research. Using Yahoo!Answers, data were extracted from the topic of diabetes and relationships between the category of “Complication and Related Disease” and other related categories were explored.

A variety of knowledge and patterns were discovered such as “acarbose might cause a side effect of hives”, “emotional depression is at onset of fibromyalgia syndrome”, “antidepressant may increase the risk of developing diabetes”, “there is a connection between imbalance and birth defects”, etc. Some of the discoveries via mining consumer generated content on social Q&A site corroborated previous findings of medical research studies (Anderberg et al. 2000; Deshauer et al. 2006; Koller and Doraiswamy 2002; Rubin et al. 2008; Wikipedia contributors 2013) and enhanced people’s understanding of these phenomena.

Future research directions on this topic include, but are not limited to: applying these visual data mining methods and technologies to reveal more patterns in other topics of interest such as depression, arthritis, asthma, etc.; employing other visual data mining methods to explore underlying knowledge about diabetes from the Q&A based social media; and exploring other social media channels with visual data mining methods.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution Noncommercial License, which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Akay A, Dragomir A, Erlandsson BE (2013) A novel data-mining approach leveraging social media to monitor and respond to outcomes of diabetes drugs and treatment. IEEE Point-of-Care Healthcare Technologies (PHT), pp 264–266
Alexa (2013) The top ranked sites in references category. http://www.alexa.com/topsites/category/Top/Reference. Accessed 24 June 2013
Anderberg UM, Marteinsdottir I, Knorring L (2000) Citalopram in patients with fibromyalgia—a randomized, double-blind, placebo-controlled study. Eur J Pain 4(1):27–35
Atzmueller M (2012) Mining social media: key players, sentiments, and communities. Wiley Interdiscip Rev Data Min Min Knowl Discov 2(5):411–419
Bian J, Topaloglu U, Yu F (2012) Towards large-scale Twitter mining for drug-related adverse events. Proceedings of the 2012 International Workshop on Smart Health and Wellbeing. Maui, Hawaii, pp 25–32
Chai K, Potdar V, Dillon T (2009) Content quality assessment related frameworks for social media. In Computational Science and Its Applications–ICCSA 2009. Springer, Berlin, pp 791–805
Comar PM, Tan PN, Jain AK (2012) A framework for joint community detection across multiple related networks. Neurocomputing 76(1S):93–104
Culotta A (2010) Towards detecting influenza epidemics by analyzing Twitter messages. 1st Workshop on Social Media Analytics (SOMA 2010). Washington DC, pp 115–122
Deshauer D, Erwin L, Karagianis J (2006) Case report: edema related to olanzapine therapy. Can Fam Physician 52(5):620
Evans BM, Kairam S, Pirolli P (2009) Do your friends make you smarter?: an analysis of social strategies in online information seeking. Inf Process Manag 46(6):679–692
Ferreira DR, Alves C (2012) Discovering user communities in large event logs. In business process management workshops. Springer, Berlin, pp 123–134
Gazan R (2011) Social QA. J Am Soc Inf Sci Technol 62(12):2301–2312
Guy S, Ratzki-Leewing A, Bahati R, Gwadry-Sridhar F (2012) Social media: a systematic review to understand the evidence and application in infodemiology. Electronic healthcare. Springer, Berlin, pp 1–8
Harper F, Raban D, Rafaeli S, Konstan J (2008) Predictors of answer quality in online QA sites. Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems. ACM, NewYork, pp 865–874
Hashimoto T, Kuboyama T, Chakraborty B, Shirota Y (2012) Discovering emerging topic about the East Japan Great Earthquake in video sharing website. In TENCON 2012–2012 IEEE Region 10 Conference. IEEE, pp 1–6
He W, Zha SH, Li L (2013) Social media competitive analysis and text mining: a case study in the pizza industry. Int J Inf Manag 33(3): 464–472
Hill S, Ready-Campbell N (2011) Expert stock picker: the wisdom of (experts in) crowds. Int J Electron Commer 15(3):73–101
Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 53(1):59–68
Kennedy H (2012) Perspectives on sentiment analysis. J Broadcast Electron Media 56(4):435–450
Kerpen D (2013) Channels. Social media marketing. http://www.smmmagazine.com/channels/. Accessed 30 Aug 2013
Kim JS, Yang MH, Hwang YI, Jeon SH, Kim KY, Jung IS et al (2012) Customer Preference Analysis Based on SNS Data. Second International Conference on cloud and Green Computing/Second International Conference on Social Computing and its Applications (CGC/SCA 2012). IEEE, pp 609–613
Kim S, Oh S (2009) Users’ relevance criteria for evaluating answers in a social QA site. J Am Soc Inf Sci Technol 60:716–727
Kim S, Oh S, Oh J (2008) Evaluating health answers in a social QA site. Proceedings of the American Society for Information Science and Technology (ASIST’08). Columbus, Ohio: Information Today, pp 1–6
Koller EA, Doraiswamy PM (2002) Olanzapine–associated diabetes mellitus. Pharmacother J Hum Pharmacol Drug Ther 22(7):841–852
Lee CH (2012) Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. Expert Syst Appl 39(18):13338–13356
Liu L, Zhu F, Zhang L, Yang S (2012) A probabilistic graphical model for topic and preference discovery on social media. Neurocomputing 95:78–88
Lobzhianidze A, Zeng W, Gentry P, Taylor A (2013) Mainstream media vs. social media for trending topic prediction—an experimental study. Consumer Communications and Networking Conference (CCNC), 2013 IEEE. IEEE, pp 729–732
Majid A, Chen L, Chen GC, Mirza HT, Hussain I, Woodward J (2013) A context-aware personalized travel recommendation system based on geotagged social media data mining. Int J Geogr Inf Sci 27(4):662–684
Robertson G, Card SK, Mackinlay JD (1989) The cognitive coprocessor architecture for interactive user interfaces. Proceedings of the Second Annual ACM SIGGRAPH Symposium on User Interface Software and Technology. ACM Press, New York, pp 10–18
Rubin RR, Ma Y, Marrero DG, Peyrot M, Barrett-Connor EL, Kahn SE, Knowler WC (2008) Elevated depression symptoms, antidepressant medicine use, and risk of developing diabetes during the diabetes prevention program. Diabetes Care 31(3):420–426
Rosenbaum H, Shachaf P (2010) A structuration approach to online communities of practice: The case of QA communities. J Am Soc Inf Sci Technol 61(9):1933–1944
Shah C, Oh JS, Oh S (2008) Exploring characteristics and effects of user participation in online social QA sites. First Monday 13(9). http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2182/2028. Accessed 15 Jan 2012
Wang M, Yang K (2012) Constructing visual tag dictionary by mining community-contributed media corpus. Neurocomputing 95:3–10
Wegrzyn-Wolska K, Bougueroua L (2012) Tweets mining for French Presidential Election. 2012 FOURTH INTERNATIONAL CONFERENCE ON COMPUTATIONAL ASPECTS OF SOCIAL NETWORKS (CASoN). IEEE, pp 138–143
White JS, Matthews JN, Stacy JL (2012 May) Coalmine: an experience in building a system for social media analytics. In SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, pp 84080A–84080A
Wikipedia contributors (2013) Atenolol. Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Atenolol. Accessed 30 July 2013
Zhang D, Si L, Rego VJ (2012) Sentiment detection with auxiliary data. Inf Retr 15(3–4SI):373–390
Zhang J (2008) Visualization for information retrieval. Springer, Berlin
Zhang J, Wolfram D, Wang P (2009) Analysis of query keywords of sports-related queries using visualization and clustering. J Am Soc Inf Sci Technol 60(8):1550–1571
Zhang J, Wolfram D (2009) Visual analysis of obesity-related query terms on Health Link. Online Information Review, 33(1), 43–57.
Zhang J, Wolfram D, Wang P, Hong Y, Gillis R (2008) Visualization of health-subject analysis based on query term co-occurrences. J Am Soc Inf Sci Technol 59(12):1933–1947
Zhang J, Zhao Y (2013) A user term visualization analysis based on a social question and answer log. Inf Process Manag 49(5):1019–1048
Zhang Y (2010) Contextualizing consumer health information searching: an analysis of questions in a social QA community. Proceedings of the 1st ACM International Health Informatics Symposium. ACM, New York, pp 210–219