Journal of Enterprise Information Management

Text mining stackoverflow: An insight into challenges and subject-related difficulties faced by computer science learners

Arash Joorabchi Michael English Abdulhussain E. Mahdi

Text mining stackoverflow

An insight into challenges and subject-related difficulties faced by computer science learners

Arash Joorabchi
Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland
Michael English
Department of Computer Science and Information Systems, University of Limerick, Limerick, Ireland, and
Abdulhussain E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland

Received 24 November 2014
Revised 9 April 2015
3 July 2015
17 July 2015
Accepted 4 August 2015

Abstract

Purpose – The use of social media and in particular community Question Answering (Q&A) websites by learners has increased significantly in recent years. The vast amounts of data posted on these sites provide an opportunity to investigate the topics under discussion and those receiving most attention. The purpose of this paper is to automatically analyse the content of a popular computer programming Q&A website, StackOverflow (SO), determine the exact topics of posted Q&As, and narrow down their categories to help determine subject difficulties of learners. By doing so, the authors have been able to rank identified topics and categories according to their frequencies, and therefore, mark the most asked about subjects and, hence, identify the most difficult and challenging topics commonly faced by learners of computer programming and software development.

Design/methodology/approach – In this work the authors have adopted a heuristic research approach combined with a text mining approach to investigate the topics and categories of Q&A posts on the SO website. Almost 186,000 Q&A posts were analysed and their categories refined using Wikipedia as a crowd-sourced classification system. After identifying and counting the occurrence frequency of all the topics and categories, their semantic relationships were established. This data were then presented as a rich graph which could be visualized using graph visualization software such as Gephi.

Findings – Reported results and corresponding discussion has given an indication that the insight gained from the process can be further refined and potentially used by instructors, teachers, and educators to pay more attention to and focus on the commonly occurring topics/subjects when designing their course material, delivery, and teaching methods.

Research limitations/implications – The proposed approach limits the scope of the analysis to a subset of Q&As which contain one or more links to Wikipedia. Therefore, developing more sophisticated text mining methods capable of analysing a larger portion of available data would improve the accuracy and generalizability of the results.

Originality/value – The application of text mining and data analytics technologies in education has created a new interdisciplinary field of research between the education and information sciences, called Educational Data Mining (EDM). The work presented in this paper falls under this field of research;

This research was funded under the "Research & Practice in ICT Learning" initiative – University of Limerick.

Journal of Enterprise Information Management
Vol. 29 No. 2, 2016
pp. 255-275

1741-0398
DOI 10.1108/JEIM-11-2014-0109

and it is an early attempt at investigating the practical applications of text mining technologies in the area of computer science (CS) education.

Keywords Text mining, Course design and delivery, Educational data mining, Social learning, Technology supported learning

Paper type Research paper

1. Introduction

The ever increasing volumes of data and information shared on social media and collaborative sites have become a rich and valuable source of knowledge for a wide spectrum of users' needs. When there is a need to learn about a new subject or to solve a specific problem, by nature people look for fast access to relevant information that would help them address that need. As such, very often they tend to consult relevant web communities, such as social media, online forums, and Community Question Answering (CQA), traditionally known as Question Answering (Q&A) sites, which gather contributions from a large array of users with different levels of expertise. Recent years, therefore, have witnessed the emergence and growing popularity of these sites among learners and educational communities, particularly students in higher education who seek to find help with their course work and material. According to Pearson's latest annual report on the use of social media for teaching and learning (Seaman and Tinti-Kane, 2013), the use of social media in higher education institutes has been on a steady rise in recent years. This includes the use of a wide range of social media websites and technologies, such as Twitter, Facebook, and Q&A sites, such as the StackExchange[1] and Quora[2].

There is currently a debate among teachers and educators with regards to the pedagogical approach of the Q&A sites to "helping" people and its productivity and contribution to real learning[3][4] and whether students should be encouraged to avail of the type of help provided by these systems[5]. This debate seems to demonstrate mixed opinions which, from our perspective, are driven by the fact that the ultimate goal of responsible teachers and educators is to facilitate the learning of their students in an active manner, help them achieve the learning outcomes and therefore the skills which their study programmes are designed for. Notwithstanding this debate, and rather than looking at the issue of students use of these Q&A sites with caution, we believe there is an opportunity here for teachers and educators to learn about and discover various aspects of this learning behaviour that can potentially help enhance their own teaching approaches. The vast amount of data and content posted on these online mediums can potentially be mined and analysed by teachers and other stakeholders to gain a detailed insight into the needs and challenges faced by the learners, and consequently used to improve the content and delivery of the courses they teach. Motivated by this opportunity, the last few years have seen the emergence of a new field of research called educational data mining (EDM) (Romero *et al.*, 2010). According to the international EDM society, "educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in".

One of the subject fields that present themselves as rich candidates for conducting EDM research is computer science (CS), which includes computer programming and software development. This is due to the fact that learners and educators who work in this field possess the IT skills required to readily utilize the services offered by Q&A sites and other social media tools, making them the most likely early adopters of social

media technologies for educational purposes (Treude *et al.*, 2012; Singer *et al.*, 2013a). This, in fact, is clearly reflected by the relatively large size of StackOverflow[6] (SO) for example, which is a social Q&A website focused on computer programming, compared to the other Q&A websites in the StackExchange network which cover over 100 various topics. Analysing the Q&As posted on SO could reveal interesting insight into the challenges faced by both novice and experienced programmers, including students. However, the sheer volume of content and information exchanged via SO and similar forums makes the task of their manual analysis difficult. Faced by this challenge, we believe a whole host of text mining and visualization methods, which have already been successfully applied in other application domains (Aggarwal and Zhai, 2012), may also be deployed to facilitate the process of analysing the content of these educational forums.

To date EDM research on CQA, and SO in particular, has mainly focused on aspects such as prediction of answers quality, crowd knowledge ranking, and user suitability (Harper *et al.*, 2008; Megan, 2014; Souza *et al.*, 2014; Treude *et al.*, 2011), user profiling and expert identification (Ginsca and Popescu, 2013; Bazelli *et al.*, 2013; Bian *et al.*, 2009), factors that contribute to the success of CQA (Mamykina *et al.*, 2011a; Liu and Agichtein, 2008), as well as other specific subject-related analyses (Tausczik *et al.*, 2014; Saxe *et al.*, 2013; Wong *et al.*, 2013; Vasilescu *et al.*, 2013). We believe, there is an apparent lack of EDM work in the context we raised above: how can teachers and educators learn from the vast amount of data produced on CQA sites to acquire an in-depth look of the subject-related needs and challenges faced by their students, utilize that knowledge to address any gaps in their teaching, delivery and coverage of the subjects, and hence enhance the learning process of their students. This paper describes a first-level attempt to addressing above research question in relation to the fields of CS, computer programming, and software development, by proposing a novel text mining approach to investigate the potential and benefits of SO to higher education instructors and teachers. The paper describes the technical details and results of a text mining method used to discover the well-defined topics and categories which have been most frequently asked about in SO. This has been achieved by adopting Wikipedia as a controlled vocabulary for detecting topics of Q&As posted on SO efficiently, where each topic corresponds to a unique Wikipedia page and the topics are clustered into categories via the Wikipedia's classification system. We also present our findings in terms of the insight gained in relation to the possible underlying reasons behind the high frequency of some topics and categories, and whether these topics and categories may be perceived to reflect subject-related difficulties faced by student-learners in these fields.

The work presented in this paper aims to address the following two research questions:

RQ1. How can text mining of CQA, and SO in particular, be utilized to provide an in-depth perception and understanding of subject-related difficulties faced by learners?

RQ2. Can instructors, teachers, and educators in higher education learn from such an insight to support their students and enhance their learning experience?

The rest of the paper is organized as follows: Section 2 introduces the StackExchange network and its most popular website, SO. Section 3 introduces text mining and describes how Wikipedia could be used as a controlled vocabulary for topic detection

and categorization of Q&As posted on SO. Section 4 presents the methodology adopted in this research and justifies this approach while Section 5 discusses the implementation details of our proposed method for text mining the Q&As. Section 6 presents the results and discusses our findings. This is followed by Section 7 which provides a conclusion along with a summary account of planned future work.

2. StackExchange and SO

StackExchange[7] is a network of Q&A websites each covering a specific topic (e.g. mathematics, physics, biology) in broad areas such as technology, science, business etc. According to Alexa[8] it currently ranks at 170 in terms of global traffic. The network currently contains 119 topic websites and 119 meta websites. Each website covering a specific topic has an accompanying meta website dedicated to its management issues. The StackExchange platform allows all users to create, vote for, and edit questions and answers and uses popularity voting as an effective mechanism for rank and filtering. It also deploys gamification and game design elements such as using rewards in the form of badges to encourage and stimulate community participation (Singer *et al.*, 2013b).

SO was the first website in the StackExchange network, which was created in 2008, and currently is the most popular website in the network. It is a free Q&A website facilitating the exchange of knowledge between both novice and experienced computer programmers. Users post and answer questions related to computer programming and may comment and rate both questions and answers. SO currently has over 3.5 million registered users. Since its inception in 2008, more than 8 million questions have been posted on the site and over 14 million answers have been provided[9], all contributing to a large knowledge repository of computer programming and software development. Parnin and Treude (2011) investigated the documentation resources that programmers use by analysing Google search results for a popular API, jQuery, and found that SO appears on the first search results page, at least once, in case of 84 per cent of the search queries. Although one might argue that this evidence only shows the popular usage of SO for discussions on a particular technology (Barua *et al.*, 2014) showed that SO covers a wide range of technologies. Currently an average of 7,000 questions are posted on the site daily and; as of August 2010, SO had an answer rate above 90 per cent and a median answer time of only 11 minutes (Mamykina *et al.*, 2011b). According to Nasehi *et al.* (2012), as of February 2012, the median time of accepted answers being posted on the site was 24 minutes and in the first hour 70 per cent of questions received their first answer. Furthermore, SO contains a complete record of every registered user including his/her badges, points, and scores which may be utilized for various educational research purposes. Before the advent of social Q&A websites, the main mechanisms for Q&A consisted of technical forums where content is available in the form of threaded discussions. The main problem with this approach is that useful information is typically mixed with irrelevant context. Social Q&A websites such as SO, on the other hand, make use of collaborative filtering to rank best answers, which are shown up front saving users' time and effort (Treude *et al.*, 2012).

3. Text mining using Wikipedia as a knowledge base

Text mining is the analysis of data contained in natural language text using techniques and tools designed to discover and extract knowledge from unstructured data (Feldman and Sanger, 2006). Text mining works by transposing words and phrases in

unstructured data into numerical values or more structured topics which can then be linked and analysed with traditional data mining techniques. Accordingly, text mining involves three major activities: information retrieval, which gathers relevant texts; information extraction, which identifies and extracts a range of specific types of information from texts of interest; and data mining, which finds associations among the pieces of information extracted from many different texts.

The data mining phase of the text mining approach can include text classification. Traditionally this has been achieved by representing the documents as bags of words and training a generic machine learning-based classifier to learn a classification model from a set of manually labelled documents. The learnt model in then used to classify unseen documents. However, such an approach to text classification ignores the important semantic relationships that may exist between keywords. This limitation can be overcome by utilizing an existing knowledge base that identifies such links between keywords and phrases. A knowledge base such as a controlled vocabulary is a way of organizing or structuring knowledge to support information retrieval. They are used in subject indexing systems such as subject headings and thesauri. Examples of such controlled vocabularies include Library of Congress Subject Headings (LCSH) and the Dewey Decimal Classification (DDC) system. In recent years, Wikipedia as a controlled vocabulary has received a lot of attention from researchers working in the field of information retrieval and knowledge management (Medelyan *et al.*, 2009). As one of the most comprehensive external knowledge sources currently available, Wikipedia has been successfully used in a wide range of applications, such as named entity recognition (Bunescu and Pasca, 2006), text classification (Wang and Domeniconi, 2008), text clustering (Hu *et al.*, 2009), event detection (Ciglan and Nørvåg, 2010), topic indexing (Medelyan *et al.*, 2008), and semantic relatedness measurement (Milne and Witten, 2008). In this work Wikipedia is used as a controlled vocabulary to support the classification of SO Q&A posts into topics and higher level categories.

Wikipedia is the world's largest free online encyclopedia. The English Wikipedia alone currently contains more than 4 million articles (Wikipedia, 2014b). Wikipedia articles are written, edited, and kept up-to-date and accurate (to a large degree) by a vast community of volunteer contributors, editors, and administrators who are collectively called Wikipedians. An investigation conducted by *Nature* (Giles, 2005) suggested that Wikipedia comes close to *Encyclopaedia Britannica* in terms of the accuracy of its science entries, although this suggestion was later disputed by Britannica (2006). However, regardless of occasional controversies around the accuracy of its articles, Wikipedia is serving a significant role in fulfilling public information needs. For example, results of a nationwide survey conducted in the USA in 2007 showed that Wikipedia attracted six times more traffic than the next closest website in the "educational and reference" category and preceded websites such as Google Scholar and Google Books with a large margin (Rainie and Tancer, 2007).

4. Methodology

As indicated in the introduction, to the best of our knowledge, the work reported here represents a new direction in the field of EDM using a novel text mining-based method to analyse Q&A websites and discover new phenomena. Hence, we have adopted a heuristic research approach combined with a new text mining approach, which utilizes Wikipedia as a knowledge base, as a research methodology to answering the research questions posed earlier. The heuristic approach is an exploratory approach to research, that is quite different from other approaches in that it is not concerned with finding

theories or testing hypotheses by following some pre-established formula, but is concerned directly with discovery of knowledge or some desired result by exploration, experimental work, intelligent analysis, and logic reasoning (Moustakas, 1990; Abbass, 2001).

The text mining approach has been adopted as an alternative to a labour-intensive manual approach such as content analysis (Krippendorff, 2012) due to the size of the data set being analysed which limits the extent to which a manual analysis approach could be adopted. The three phases of the text mining approach are now described with an overview provided in Figure 1 and further technical detail described in the Section 5.

Information retrieval (data collection): The data for this study has been extracted for the SO website. SO currently has over 3.5 million registered users. Since its inception in 2008, more than 8 million questions have been posted on the site and over 14 million answers have been provided (see footnote 9). Approximately 7,000 questions are posted on a daily basis. Thus within the software development field, SO is the largest community question and answer forum and therefore is a rich data source for the study proposed in this wok. In addition, the StackExchange network, of which the SO website is part, has adopted an open data policy whereby all of its websites' data such as posts, users, votes, and comments are made freely available to the public. An anonymized dump of all the user-contributed StackExchange data are published periodically[10]. Therefore all of the SO data related to questions, answers, and users is freely available for download and analysis.

Given that Wikipedia is being used as a controlled vocabulary, only the questions and answers that contain links to Wikipedia pages are analysed in this study. This phase of the text mining approach involves the identification of all the Q&A posts on SO that may contain a link to a valid Wikipedia webpage. Many of the questions and answers posted on SO contain links to Wikipedia. This phase of the analysis returned a total of 188,548 matches including a combination of questions and answers.

Information extraction (data cleaning): This phase of the mining process involved extracting the actual URLs of the Wikipedia pages and validating them. There were two levels of validation involved. The first phase removed links which contained invalid characters and the second phase ensured that the syntactically correct links pointed to an actual Wikipedia URL at the time of analysis.

Data mining (visualization and frequency analysis): Each valid Wikipedia URL identified in the previous phase points to an article within the Wikipedia where the topic can have one or more parent categories. A software application was built to

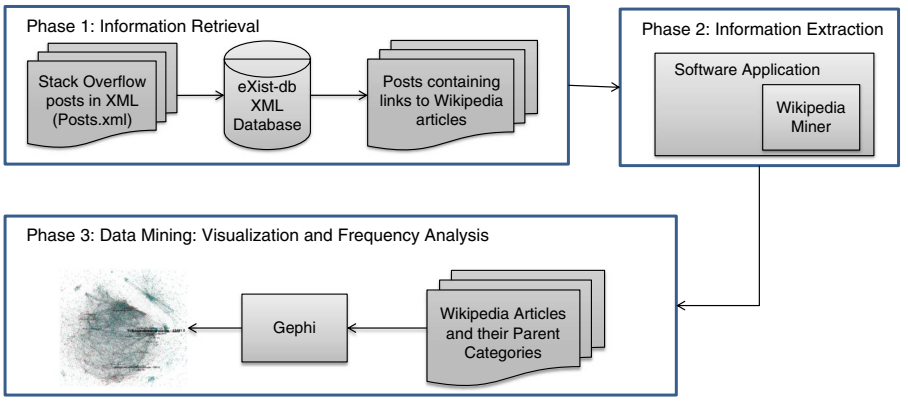

Figure 1.
Overview of text mining approach

extract the corresponding articles/topics for the URLs, their appearance frequency in the analysed Q&As, and their parent categories to facilitate the visualization and analysis presented in Section 6.

4.1 Wikipedia as a knowledge base

The SO Q&A posts that contain links to Wikipedia form the data set for analysis in this study. While this limits our analysis to a subset of SO Q&A posts, we believe that the benefits offered by using Wikipedia as a controlled vocabulary outweigh the limitations of focusing on a subset of the Q&A posts. As a controlled vocabulary, Wikipedia offers a number of advantages over similar controlled vocabularies such as library classification systems:

(1) Extensive coverage and comprehensiveness: the English Wikipedia currently contains over 4 million articles covering subjects in all aspects of human knowledge.

(2) Up-to-dateness: due to the crowd-sourced nature of Wikipedia and its large pool of editors, Wikipedia articles are generally well-maintained and kept quite up-to-date. For example, a study examining the potential of combining Twitter and Wikipedia data for event detection showed that in case of major events Wikipedia lags Twitter only by about three hours (Osborne *et al.*, 2012).

(3) Rich description: Wikipedia articles provide rich descriptive content for the represented concepts describing their various aspects in details.

(4) Multilingual: as of July 2014 Wikipedia exists in more than 287 languages. Wikipedia has more than 1 million articles in each of the 12 most populated languages and more than one hundred thousand articles in each of the 52 less populated languages (Wikipedia, 2014a).

According to part 1 of the international standard for thesauri (ISO 25964-1)[11], a compliant thesaurus should capture and encode three main types of relationship between concepts: equivalence relations between synonyms and near-synonyms, e.g. car and automobile; hierarchical relations between broader and narrower concepts, e.g. vehicle and car; associative relations between concepts that are closely related in a non-hierarchical fashion, e.g. Formula 1 and car. Adapted as a controlled vocabulary, Wikipedia meets all these requirements: each Wikipedia article has a descriptor which is the preferred and most commonly used term for the represented concept, and each article is assigned a set of non-descriptors which are the less commonly used synonyms and alternative lexical forms for the concept (i.e. equivalence relations); similar to the notion of "Related Terms" in traditional controlled vocabularies, related articles in Wikipedia are connected via hyperlinks (i.e. associative relations); each Wikipedia article is classified according to the Wikipedia's own community-built classification scheme into one or more broader categories, which resembles the notion of "Broader Terms/Narrower Terms" in traditional controlled vocabularies (i.e. hierarchical relations).

While the accuracy of some Wikipedia articles has been disputed as discussed in Section 3, and the editorial quality of some articles is poor, this does not seem to effect the quality of the controlled vocabularies derived from Wikipedia. For example (Milne *et al.*, 2006), investigated the application of Wikipedia as a thesaurus in the domain of agriculture and compared it with a manually created professional thesaurus in this domain, Agrovoc, as the gold standard. They found that Wikipedia contains a substantial proportion of concepts and semantic relations encoded in Agrovoc and has

impressive coverage of contemporary documents in the domain. In a similar study (Vivaldi and Rodríguez, 2010) derived three domain-specific thesauri for astronomy, chemistry, and medicine in two languages (English, Spanish) from Wikipedia, and reported promising results in terms of the coverage and accuracy of the constructed thesauri. These and a number of similar studies have shown that Wikipedia is an effective source of knowledge for constructing various types of controlled vocabularies, including thesauri, taxonomies, and ontologies (Ponzetto and Strube, 2007; Fogarolli, 2011; Milne *et al.*, 2007). Traditional controlled vocabularies such as LCSH and DDC have been created to support the indexing and classification of books and are only updated periodically. Alternatively, Wikipedia as a crowd-sourced classification system is updated in real-time and thus is likely to include new and emerging topics and categories associated with the computer programming and software development domains.

The above listed advantages of Wikipedia over other controlled vocabularies and more importantly the existing links between SO posts and Wikipedia articles make Wikipedia our controlled vocabulary of choice for identifying the topics and categories of the SO Q&As in our proposed method.

5. Implementation of text mining process

This section discusses the technical details involved in implementing the text mining process and subsequent analysis. This involves the following four main steps:

(1) develop a text mining method, which utilizes Wikipedia as a controlled vocabulary for determining the topics and categories of SO Q&A posts;

(2) apply developed text mining method to the most recent version of SO dump to categorize and cluster Q&A posts, producing an interactive visual representation of resulting topics and categories;

(3) using the output from above to closely inspect resulting topics and categories of the contents of Q&A posts in SO in terms of their frequencies, rankings, and associations; and

(4) analyse the findings from (3) to gain insight into possible reasons behind the high frequency of some topics and categories, and whether these topics and categories reflect subject-related difficulties faced by student-learners in these fields, in order to infer how best our text mining method of SO can be used to inform instructor and teachers in designing the coverage, delivery, and teaching approach of their CS courses.

The rest of this section details the first two steps outlined above. The "Results & Discussion" section, on the other hand, incorporates description of steps (3) and (4). The text mining method we have developed and applied in this work to SO posts, to determine and categorize their topics, involves the following four processes: uploading the SO posts into a database; querying the database and retrieving the posts which contain a link to Wikipedia; finding the parent categories of cited Wikipedia articles; and building the graph of cited articles and their parent categories.

5.1 Uploading the data into a database

Each site in the StackExchange network of which SO is part, is formatted as a separate archive consisting of zipped XML data files for Posts, Users, Tags, Votes, Comments, Badges, PostHistory, and PostLinks (for complete schema information, see the

readme file[12]). In order to easily access and analyse the SO posts, we created an XML database using the eXist-db (Meier, 2014), which is an open-source native XML database engine, to store the SO data in. We then downloaded the most recent version of the SO dump and uploaded the Posts.xml file which contains all the posted questions and answers up-to-date into the database. This enables us to search and retrieve posts of interest by querying the database using the XPath and XQuery languages.

5.2 Querying the database

Each single post inside the Posts.xml file has multiple attributes including: a unique identifier, a PostTypeId ("1" for questions and "2" for answers), Title, and Body. We utilize these attributes to retrieve those posts which their content/body is of interest to us. In order to examine and analyse the SO posts in terms of their topics and categories, we have used Wikipedia as a crowd-sourced classification system. We used the following XPath query to retrieve all the Q&A posts which included a link to the Wikipedia articles:

```
for $x in doc("/db/SO/Posts.xml")/posts/row[contains(@Body,"http://en.wikipedia.
org/wiki")]
return $x
```

This query returned a total of 188,548 matches, of which 20,924 were questions, 165,080 were answers, and 2,544 were of miscellaneous type, e.g., moderator nomination. After removing miscellaneous posts and those with invalid Wikipedia URLs (i.e. those containing restricted characters[13]), the remaining total of 186,004 Q&A posts contained 28,648 unique URLs linking to the Wikipedia articles with a total appearance frequency of 230,887. Finally after confirming the validity of all the discovered unique links by retrieving their corresponding pages on the Wikipedia website, a total of 21,366 were verified as valid links to existing articles on Wikipedia.

5.3 Finding the parent categories

We have deployed an open-source toolkit called Wikipedia-Miner (Milne, 2009) to discover the parent categories of the Wikipedia articles which were linked to in the SO posts. Wikipedia-Miner effectively unlocks Wikipedia as a general-purpose knowledge source for natural language processing (NLP) applications by providing rich semantic information on topics and their lexical representations beyond that offered by domain-specific thesauri. Utilizing the Java programing language and the Wikipedia-miner environment, we first converted the Wikipedia URLs identified in the Q&A posts into article objects using the getArticleByTitle function of the Wikipedia-miner API. We then used the getParentCategories function to retrieve the parent categories of each article.

5.4 Data mining: visualization and frequency analysis

Finally, we used the collected information to build a directed graph representing articles and categories as nodes and their parent-child relationships as edges. The graph was encoded and stored as an xml file in GraphML[14] format which may be easily imported into popular open-source network analysis and visualization software packages such as the Gephi platform (Bastian *et al.*, 2009) used in this work. Gephi is an interactive visualization platform to explore and understand graphs. It enables users to interact with the representation, manipulate the structures, shapes, and colours to reveal hidden properties and discover patterns in graphs. Utilizing Gephi enables us to visually analyse the graph of Wikipedia article citations in SO posts and examine their usage patterns to find the topics and subjects which are most asked about.

6. Results and discussion

In this section we discuss our findings that resulted from the analysis of the graph of topics and categories, which was created by mining the content of Q&As in SO. The Gephi project file containing the graph of Wikipedia articles cited in SO posts and their frequency and relationship details is available online[15] and may be readily used for further studies. Figure 2 shows a zoomed-out overview of the graph. This graph contains all the 21,366 unique Wikipedia articles mentioned in SO, as well astheir 20,587 direct parent categories.

Since May 2004, Wikipedia authors have been categorizing Wikipedia articles according to a community-built classification system (a.k.a. folksonomy) which has been growing rapidly. Including the parent categories of articles in the graph allows us to cluster-related articles together and draw more accurate conclusions regarding the dominant subjects being discussed. The nodes in the graph are colour coded to differentiate their type, i.e., red for articles and green for categories. Also, the node sizes are set to reflect the appearance frequency of the articles and categories that they represent. The resulting graph has a total of 41,923 nodes and 71,160 edges. Wikipedia articles and categories are classified according to a hierarchical classification system with (currently) 16 levels of depth reflecting speciality/generality. In the rest of this section we discuss the articles and categories at various

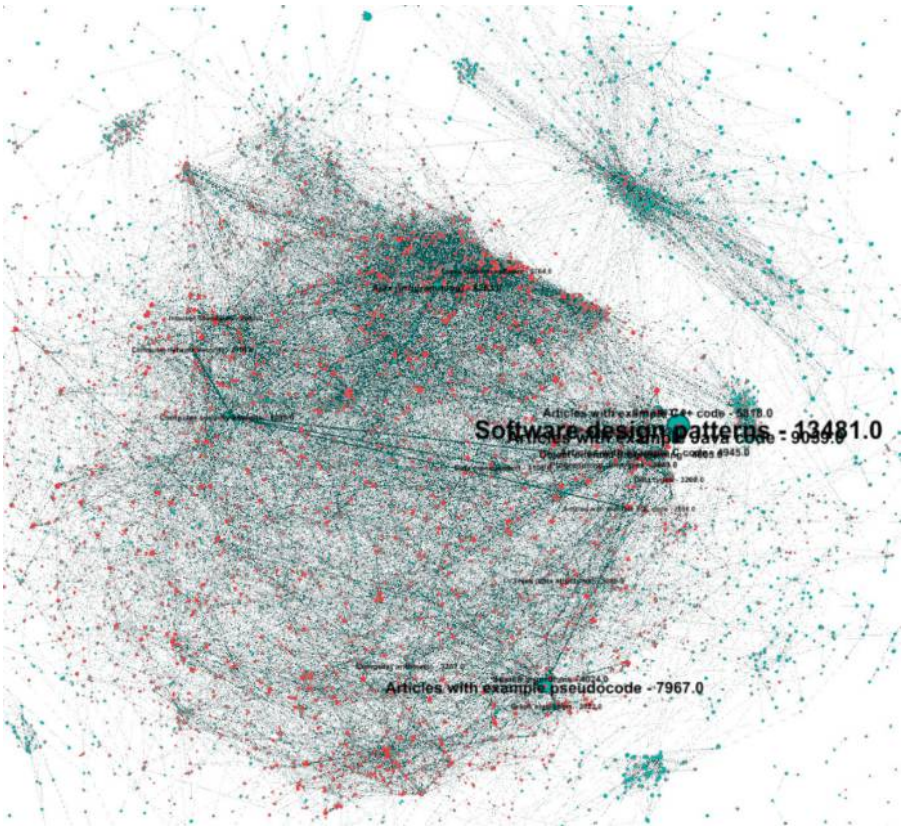

Figure 2.
An overview of the StackOverflow-Wikipedia graph

levels of the hierarchy which have been most frequently mentioned in SO posts. Table I lists the top 50 most prominent articles along with their depth and frequency. As evident from these results most high-frequency articles belong to levels 4-8 of the Wikipedia classification system, where the topics are neither too generic nor too detailed. The most frequently mentioned article, "Same origin policy" discusses an important security concept for a number of browser-side programming languages, such as javascript. Figure 3 shows a zoom-in view of this paper along with its parent category and sibling nodes in the graph and their corresponding frequencies in the Q&A posts.

The list in Table I contains a substantial number of articles covering various aspects of software engineering such as "Singleton pattern", "Model-view-controller", "Dependency injection", and "Strategy pattern". Specifically, articles related to the object-oriented programming paradigm and to software design and development have an overwhelming presence. These include: "Factory method pattern", "Observer pattern", "Single responsibility principle", "Visitor pattern", "Decorator pattern", "Command pattern", "Resource Acquisition Is Initialization", "Liskov substitution principle", and "Object-relational mapping". The remaining articles fall into a number of minor groups covering various aspects of CS including:

(1) databases: articles covering relational databases and their query languages including: "SQL injection", "Database normalization", "Entity-attribute-value model", and "Join (SQL)";

(2) web technologies: articles covering a wide range of Web technologies including: "JSONP", "JSON", "Representational state transfer", "Ajax (programming)", "Cross-site scripting", "GET (HTTP)", "UTF-8", "Data URI scheme", "Cross-origin resource sharing", "Unobtrusive JavaScript", and "Hypertext Transfer Protocol"; and

(3) data structures and algorithms: articles covering various data structures and related algorithms including: "Trie", "Levenshtein distance", "Hash table", "Fisher-Yates shuffle", and "Dijkstra's algorithm".

In the next step of the analysis, we aggregated the weights (i.e. appearance frequencies) of all the articles which shared the same parent category(ies) and assigned these total weights to the corresponding parent categories. This process in practice acts as a clustering technique where all related articles are detected and their weights are combined to provide us with a different vantage point on the data and the popular categories is SO. Table II provides a list of the top 50 categories with highest frequencies in SO posts. As shown in Table II the most frequent category is "Software design patterns" which encompasses all the high-frequency articles in Table I covering various design pattern concepts in software engineering. Categories covering sample codes in Java, pseudocode, C++, and C have the second highest frequencies. This highlights the fact that questions related to the Java programming language are the most common in SO followed by those related to C++. The category "Object-Oriented Programming" is ranked 6th which confirms the popularity of Q&As addressing OOP concepts in SO. Search, graph, and sorting algorithms are ranked 9th, 16th, and 26th, respectively. These algorithms and their variations are covered in most undergraduate CS courses, and therefore we suspect most of these questions are posted by CS students and novice programmers. The remaining top categories cover a wide range of topics in CS from programming principles to formal languages.

Article title	Appearance frequency	Hierarchical depth
Same-origin policy	2,106	5
SQL injection	1,536	5
Model view controller	1,072	5
Singleton pattern	988	5
Comet (programming)	829	6
JSONP	795	7
Factory method pattern	788	5
Trie	781	7
Database normalization	761	5
Dependency injection	751	5
Observer pattern	744	5
Single responsibility principle	718	4
Levenshtein distance	707	5
Strategy pattern	652	5
Do not repeat yourself	644	4
JSON	629	6
Representational state transfer	626	6
Byte order mark	623	8
Base64	595	6
Ajax (programming)	594	6
Cross-site scripting	578	7
GET (HTTP)	568	4
UTF-8	561	5
Visitor pattern	543	5
Data URI scheme	541	6
Resource acquisition is initialization	538	7
Join (SQL)	524	7
Floating point	522	6
Endianness	522	3
Decorator pattern	518	5
Liskov substitution principle	511	4
Hash table	504	6
Cron	496	7
Regular expression	487	6
Fisher-Yates shuffle	486	6
ISO 8601	484	4
Two's complement	472	7
IEEE floating point	469	7
Short-circuit evaluation	448	7
Cross-origin resource sharing	431	7
Command pattern	422	5
Object-relational mapping	420	6
Dijkstra's algorithm	417	6
C++11	413	7
Unobtrusive JavaScript	400	5
Big O notation	393	5
Entity attribute value model	382	6
Unix time	375	4
Hypertext transfer protocol	373	6

Table I.
Top 50 Wikipedia articles with highest appearance frequencies in SO posts

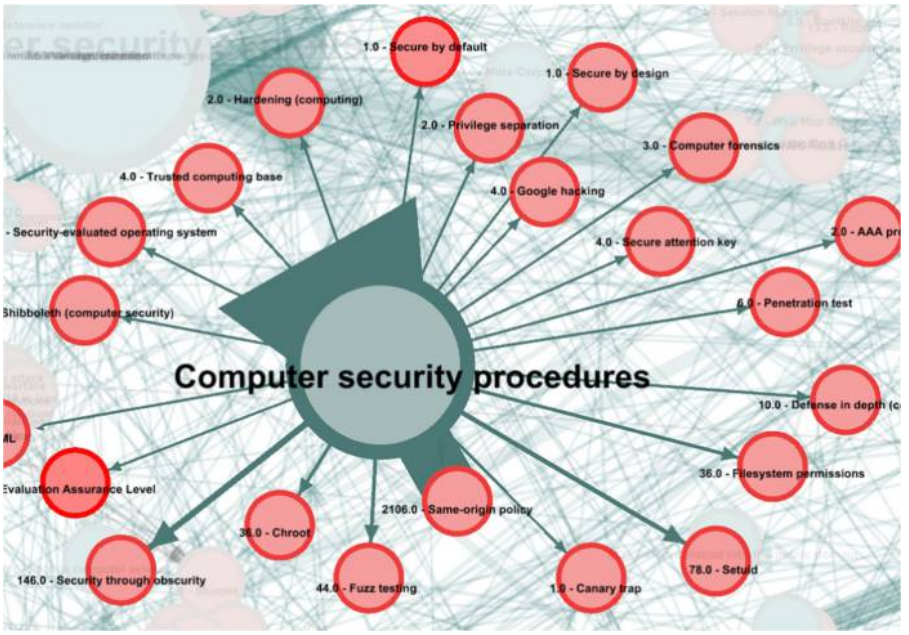

Figure 3.
A zoom-in view of the article "Same origin policy" and its parent and sibling nodes in the graph

The data presented in Tables I and II may be used to prioritize the topics covered in undergraduate and graduate CS courses and guide the redesign and update of the courses and their materials. Although this data provides an up-to-date and accurate view of topics of interest frequently asked about and discussed in SO, conducting further detailed manual analysis of these topics could shed more light on the reasons behind their high frequency. For example, the data presented in Table I has indicated that questions related to the topic of "regular expressions" have a high frequency in SO. However, it does not give us any further clues as to what aspect of regular expressions those Q&As cover, and therefore, still some level of manual analysis is required. To demonstrate this, we have conducted further analysis on a random subset of 50 Q&As discussing the topic of regular expressions. Looking at the Q&As citing this particular article in Wikipedia, it turns out that the majority of questions posted in this regard (39/50) evolve around the formulation of regular expressions to match specific patterns of interest. For example, the title of some of these questions read:

- how to validate if student number is a 11-digit number;
- regular expression to match only alphabetic characters;
- please explain this e-mail validation regular expression; and
- remove everything except what is in quotes java.

The remaining questions regarding the topic of regular expressions relate to its other aspects such as deployment in various environments and programming languages. Examples of these questions include:

- what does this split line in Scala mean?
- JavaScript – How to get at specific value in a string?

Category title	Combined weight	Hierarchical depth
Software design patterns	13,481	4
Articles with example Java code	9,059	6
Articles with example pseudocode	7,967	6
Articles with example C++ code	5,818	6
Articles with example C code	4,945	6
Object-oriented programming	4,605	7
Ajax (programming)	4,263	7
C++	4,037	7
Search algorithms	4,024	5
Programming principles	3,449	3
Software architecture	3,428	5
Computer arithmetic	3,257	6
Computer security exploits	3,255	6
Data types	3,208	5
Data management	3,108	4
Graph algorithms	3,022	6
Trees (data structures)	3,000	6
Internet standards	2,867	6
Computer network security	2,794	6
Cross-platform software	2,764	8
Articles with example SQL code	2,666	6
Computer security procedures	2,495	4
Application layer protocols	2,472	6
JavaScript	2,470	7
Data serialization formats	2,445	5
World Wide Web Consortium standards	2,437	6
Sorting algorithms	2,408	5
Web development	2,368	6
Hypertext transfer protocol	2,357	6
Java platform	2,321	9
Articles with example C Sharp code	2,303	6
Dynamic programming	2,291	5
Type theory	2,265	4
SQL	2,259	6
Injection exploits	2,221	7
Binary arithmetic	2,197	6
Microsoft application programming interfaces	2,136	8
Method (computer programming)	2,135	8
Programming paradigms	2,110	6
Web 2.0 neologisms	2,100	5
C programming language	2,096	8
Subroutines	1,986	7
Scripting languages	1,979	6
NET framework	1,959	6
Software development philosophies	1,955	6
Articles with example Python code	1,885	6
Markup languages	1,877	5
Formal languages	1,855	5
XML-based standards	1,794	5

Table II.
Top 50 Wikipedia categories with highest appearance frequencies in SO posts

- standard regex vs python regex discrepancy; and
- validations in textbox in vb.net.

Based on above observation, we divided the regular expressions related Q&As into three major groups "formulation", "usage", and "miscellaneous". Table III shows the title of analysed Q&As along with their grouping information. An excel file containing the data presented in Table III with active hyperlinks to the original Q&As is available online[15].

A similar type of analysis may be applied to the other high-frequency Wikipedia articles and categories of interest listed in Tables I and II to discover more detailed information in respect to the type of questions posted in each topic.

7. Conclusion

In this work we proposed a text mining method for discovering the most frequent topics and categories commonly discussed in Q&A websites. We applied the proposed approach to the largest Q&A website in the field of CS, StackOverflow.com, and presented the obtained results in terms of the most popular Wikipedia articles and categories mentioned in SO posts. This then enabled us to highlight the most frequently asked about topics and subjects in computer programming and, hence, identify difficulties faced by learners in this field.

7.1 Implications to theory and practice

Utilizing these findings, we inferred the following: the proposed text mining approach has the potential for informing instructors and teachers in higher education of the important and challenging subjects in the courses they deliver, so that they could focus more on these topics/subjects in the coverage, delivery, and teaching/training approaches of their course material. The application of data mining and data analytics technologies in education has created a new interdisciplinary field of research spanning the education and information sciences, called EDM. The work presented in this paper is an early attempt at investigating the practical application of EDM in the area of CS education. Thus, in our opinion, the contribution of this work is twofold by:

(1) demonstrating a novel application of text mining to SO and its ability in providing an in-depth analysis of subject-related difficulties and challenges in learning new topics in the field of computer programming; and

(2) identifying potential benefits of analysing the content of SO to instructors and teachers in terms of revising the contents of their computer programming related courses and the way they are delivered, such that more attention is paid to the difficult topics identified.

In practice the presented method and resulting data analysis act as an important step towards a more detailed investigation which would involve further distilling, analysis, and inference of the educational content available on Q&A websites. The proposed method may be easily applied to online forums and Q&A websites in other fields of science and engineering, such as math.stackexchange.com which is another Q&A website in the StackExchange network focusing on the field of mathematics and containing more than half a million posts.

No.	Q&A title	Type Form	Usage	Mis.
1	A: if textBox1 contains integer	✓		
2	A: regular expression with pipe	✓		
3	A: what does this split line in Scala mean?		✓	
4	A: Django Urlpatter for string	✓		
5	A: how to automatically insert pragmas in your programme	✓		
6	A: what does authorize(users = "*") mean?	✓		
7	A: how to validate if student number is a 11-digit number	✓		
8	A: VS 2012 find and replace text outside string literals	✓		
9	A: why are regular expressions called "regular" expressions?			✓
10	A: regex – getting last occurrence not first – why?	✓		
11	A: find string and replace line in Linux	✓		
12	A: who defines regular expressions?			✓
13	A: remove everything except what is in quotes java	✓		
14	A: working with conditions based on command output		✓	
15	A: having trouble creating a regular expression	✓		
16	A: please explain this e-mail validation regular expression	✓		
17	A: how can I remove "Page ###" from a string in php	✓		
18	A: regular expressions: how to express\wwithout underscore	✓		
19	A: validations in textbox in vb.net		✓	
20	A: regex for nested values	✓		
21	A: how to use string.split( ) for the following string in javascript		✓	
22	A: regex to remove junk from a .txt file in Unix	✓		
23	A: delete all lines beginning with a # from a file	✓		
24	A: regular expression to match only alphabetic characters	✓		
25	A: need a regular expression to allow only one character in a textbox in asp.net	✓		
26	A: algorithm to parse string with dictionary	✓		
27	A: regular Expression with foreign languages	✓		
28	A: preg_match return longest match	✓		
29	A: searching a folder for parts of a file's name	✓		
30	Q: use sed to search and replace patterns via regular expressions	✓		
31	A: find all occurrences between tag	✓		
32	A: how can I perform a diff that ignores all comments?	✓		
33	A: Bash RexEx: read file line by line to pull out each href in captured groups	✓		
34	A: how to validate math formular string using regex?	✓		
35	A: can a regular expression be tested to see if it reduces to *	✓		
36	A: replacing special characters by null	✓		
37	A: difference between matches and equalsIgnoreCase or equals in string class		✓	
38	A: extract substring from lines using grep, awk, sed or etc.	✓		
39	A: comparing similar words and phrases	✓		
40	A: how to write the regex for this expression	✓		
41	A: preg_replace in PHP		✓	
42	A: JavaScript – how to get at specific value in a string?		✓	
43	A: string to float and regex	✓		
44	A: how to configure URLs using PHP		✓	
45	A: grouping text into sections algorithm	✓		
46	A: grep command to extract e-mail addresses	✓		
47	A: JavaScript regex to extract text and number	✓		
48	A: help with capturing URL fragment in Django	✓		
49	A: php regular expression for video swf	✓		
50	Q: standard regex vs python regex discrepancy		✓	
Sum		39	9	2

Table III.
A sample of 50 regular expressions related Q&As categorized according to their types

7.2 Limitations

Having demonstrated the benefits and potential of the proposed text mining method reported in this paper, we would also like to highlight some of its limitations. The proposed method in its current form limits the scope of the analysis to a subset of Q&As which contain one or more links to Wikipedia. The results presented in this work are based on the analysis of a total of approximately 186,000 Q&As. We expect this sample size to be of sufficient magnitude for drawing generalized conclusions in terms of the topic and categories which the learners have the most difficulty with. However, developing more sophisticated text mining methods capable of analysing a larger portion of data could improve the accuracy and generalizability of the results.

7.3 Future work

As future work, we plan to enhance the depth and coverage of our analysis method by: attempting to automatically classify all the Q&A posts according to the Wikipedia classification system using a machine learning-based classification method, which would result in a significant increase in the sample size and subsequently the accuracy of the results; further refine the results of the analysis and its granularity by analysing the textual content of the Q&As in more depth and extracting the terms and phrases which are statistically significant based on their appearance frequencies; and investigating the application of the proposed method in analysing other fields of science and engineering, such as mathematics, and their respective online forums and Q&A websites.

Notes

1. http://stackexchange.com/
2. www.quora.com/
3. http://michael.richter.name/blogs/why-i-no-longer-contribute-to-stackoverflow
4. http://meta.stackexchange.com/questions/10811/how-do-i-ask-and-answer-homework-questions
5. http://meta.stackoverflow.com/questions/254433/should-i-send-students-to-stack-overflow
6. http://stackoverflow.com/
7. http://stackexchange.com/
8. www.alexa.com/siteinfo/stackexchange.com
9. http://stackexchange.com/sites?view=list#traffic
10. https://archive.org/details/stackexchange
11. www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53657
12. https://archive.org/download/stackexchange/readme.txt
13. https://en.wikipedia.org/wiki/Wikipedia:Naming_conventions_(technical_restrictions)
14. http://graphml.graphdrawing.org/
15. www.skynet.ie/~arash/zip/StackOverflow_Wikipedia_Sep2013.zip

References

Abbass, H.A. (2001), *Heuristic and Optimization for Knowledge Discovery*, Idea Group Pub, Hershey, PA.

Aggarwal, C.C. and Zhai, C. (2012), *Mining Text Data*, Springer-Verlag, New York, NY.

Barua, A., Thomas, S. and Hassan, A. (2014), "What are developers talking about? An analysis of topics and trends in stack overflow", *Empirical Software Engineering*, Vol. 19 No. 3, pp. 619-654.

Bastian, M., Heymann, S. and Jacomy, M. (2009), "Gephi: an open source software for exploring and manipulating networks", in Cohen, W.W., Nicolov, N., Glance, N., Hurst, M., Soboroff, I., Java, A. and Marlow, C. (Eds), *Third International AAAI Conference on Weblogs and Social Media (ICWSM09), AAAI, San Jose, CA, 17-20 May*.

Bazelli, B., Hindle, A. and Stroulia, E. (2013), "On the personality traits of stackoverflow users", *29th IEEE International Conference on Software Maintenance (ICSM), Eindhoven, 22-28 September*, pp. 460-463.

Bian, J., Liu, Y., Zhou, D., Agichtein, E. and Zha, H. (2009), "Learning to recognize reliable users and content in social media with coupled mutual reinforcement", in Quemada, J., León, G., Maarek, Y. and Nejdl, W. (Eds), *Proceedings of the 18th International Conference on World Wide Web, ACM, Madrid, 20-24 April*, pp. 51-60.

Britannica (2006), "Fatally Flawed – refuting the recent study on encyclopedic accuracy by the journal nature", Encyclopædia Britannica Inc., available at: http://corporate.britannica.com/britannica_nature_response.pdf (accessed January 2016).

Bunescu, R.C. and Pasca, M. (2006), "Using encyclopedic knowledge for named entity disambiguation", in Keller, F. and Proszeky, G. (Eds), *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, 3-7 April*, pp. 9-16.

Ciglan, M. and Nørvåg, K. (2010), "WikiPop: personalized event detection system based on Wikipedia page view statistics", in Huang, J., Koudas, N., Jones, G., Wu, X., Collins-Thompson, K. and An, A. (Eds), *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, Toronto, ON, 26-30 October*, pp. 1931-1932.

Feldman, R. and Sanger, J. (2006), *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, New York, NY.

Fogarolli, A. (2011), "Wikipedia as a source of ontological knowledge: state of the art and application", in Caballé, S., Xhafa, F. and Abraham, A. (Eds), *Intelligent Networking, Collaborative Systems and Applications*, Springer, Berlin, Heidelberg, pp. 1-26.

Giles, J. (2005), "Internet encyclopaedias go head to head", *Nature*, Vol. 438, pp. 900-901, available at: www.nature.com/nature/journal/v438/n7070/full/438900a.html

Ginsca, A.L. and Popescu, A. (2013), "User profiling for answer quality assessment in Q&A communities", in Mahmud, J., Caverlee, J., Nichols, J., O'Donovan, J. and Zhou, M. (Eds), *Proceedings of the 2013 Workshop on Data-Driven User Behavioral Modelling and Mining from Social Media, ACM, San Francisco, CA, 27 October-1 November*.

Harper, F.M., Raban, D., Rafaeli, S. and Konstan, J.A. (2008), "Predictors of answer quality in online Q&A sites", in Czerwinski, M., Lund, A. and Tan, D. (Eds), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Florence, 5-10 April*, pp. 865-874.

Hu, X., Zhang, X., Lu, C., Park, E.K. and Zhou, X. (2009), "Exploiting Wikipedia as external knowledge for document clustering", *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Paris.

Krippendorff, K. (2012), *Content Analysis: An Introduction to Its Methodology*, SAGE Publications, Thousand Oaks, CA.

Liu, Y. and Agichtein, E. (2008), "You've got answers: towards personalized models for predicting success in community question answering", *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, Association for Computational Linguistics, Columbus, OH.

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G. and Hartmann, B. (2011a), "Design lessons from the fastest Q&A site in the west", in Tan, D., Fitzpatrick, G., Gutwin, C., Begole, B. and Kellogg, W.A. (Eds), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Vancouver, BC, 7-12 May*, pp. 2857-2866.

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G. and Hartmann, B. (2011b), "Design lessons from the fastest Q&A site in the west", in Tan, D., Fitzpatrick, G., Gutwin, C., Begole, B. and Kellogg, W.A. (Eds), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Vancouver, BC, 7-12 May*, pp. 2857-2866.

Medelyan, O., Milne, D., Legg, C. and Witten, I.H. (2009), "Mining meaning from Wikipedia", *International Journal of Human-Computer Studies*, Vol. 67 No. 9, pp. 716-754.

Medelyan, O., Witten, I.H. and Milne, D. (2008), "Topic Indexing With Wikipedia", in Bunescu, R., Gabrilovich, E. and Mihalcea, R. (Eds), *First AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08), AAAI Press, Chicago, IL, 13-14 July*, pp. 19-24.

Megan, S. (2014), "A bit of code", in Christian, F. (Ed.), *How the Stack Overflow Community Creates Quality Postings*, IEEE Computer Society Conference Publishing Services (CPS), Washington, DC, pp. 1425-1434.

Meier, W. (2014), "eXist-DB, exist-db.org, released under the open source GPL licence", available at: http://exist.sourceforge.net/ (accessed January 2016).

Milne, D. and Witten, I.H. (2008), "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links", First AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08), Chicago, IL.

Milne, D. (2009), "An open-source toolkit for mining Wikipedia", *New Zealand Computer Science Research Student Conference, Auckland, 6-9 April*.

Milne, D., Medelyan, O. and Witten, I.H. (2006), "Mining domain-specific thesauri from Wikipedia: a case study", in Nishida, T., Shi, Z., Visser, U., Wu, X., Liu, J., Wah, B., Cheung, W. and Cheung, Y.-M. (Eds), *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Hong Kong, 18-22 December*, pp. 442-448.

Milne, D.N., Witten, I.H. and Nichols, D.M. (2007), "A knowledge-based search engine powered by Wikipedia", in Laender, A.H.F., Falcão, A.O., Olsen, Ø.H., Silva, M.J., Baeza-Yates, R., McGuinness, D.L. and Olstad, B. (Eds), *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ACM, Lisbon, 6-10 November*, pp. 445-454.

Moustakas, C. (1990), *Heuristic Research: Design, Methodology, and Applications*, SAGE Publications, Thousand Oaks, CA.

Nasehi, S.M., Sillito, J., Maurer, F. and Burns, C. (2012), "What makes a good code example?: a study of programming Q&A in stackoverflow. software maintenance (ICSM)", *28th IEEE International Conference*, *23-28 September*, pp. 25-34.

Osborne, M., Petrovic, S., Mccreadie, R., Macdonald, C. and Ounis, I. (2012), "Bieber no more: first story detection using Twitter and Wikipedia", *SIGIR Workshop in Time-Aware Information Access (TAIA'12)*, ACM, Portland, OR.

Parnin, C. and Treude, C. (2011), "Measuring API documentation on the web", in Treude, C., Storey, M.-A., van Deursen, A., Begel, A. and Black, S. (Eds), *Proceedings of the 2nd International Workshop on Web 2.0 for Software Engineering, ACM, Waikiki, Honolulu, HI, 21-28 May*, pp. 25-30.

Ponzetto, S.P. and Strube, M. (2007), "Deriving a large scale taxonomy from Wikipedia", in Cohn, A. (Ed.), *Proceedings of the 22nd National Conference on Artificial Intelligence, Vol 2, AAAI Press, Vancouver, 22-26 July*, pp. 1440-1445.

Rainie, L. and Tancer, B. (2007), "Wikipedia users", Pew Internet and American Life Project, available at: www.pewinternet.org/Reports/2007/Wikipedia-users.aspx (accessed January 2016).

Romero, C., Ventura, S., Pechenizkiy, M. and Baker, R.S.J. (2010), *Handbook of Educational Data Mining*, CRC Press, Boca Raton, FL.

Saxe, J., Mentis, D. and Greamo, C. (2013), "Mining web technical discussions to identify malware capabilities", *Distributed Computing Systems Workshops (ICDCSW), 2013 IEEE 33rd International Conference*, *8-11 July*, pp. 1-5.

Seaman, J. and Tinti-Kane, H. (2013), "Social media for teaching and learning", Pearson Learning Solutions, Boston, MA, available at: www.pearsonlearningsolutions.com/higher-education/social-media-survey.php

Singer, L., Figueira Filho, F. and Storey, M.-A. (2013a), "Software engineering at the speed of light: how developers stay current using Twitter", Technical Report No. DCS-350-IR, University of Victoria, Victoria.

Singer, L., Filho, F.F., Cleary, B., Treude, C., Storey, M.-A. and Schneider, K. (2013b), "Mutual assessment in the social programmer ecosystem: an empirical investigation of developer profile aggregators", in Bruckman, A., Counts, S., Lampe, C. and Terveen, L. (Eds), *Proceedings of the 2013 Conference on Computer Supported Cooperative Work, ACM, San Antonio, TX, 23-27 February*, pp. 103-116.

Souza, L.B.L.D., Campos, E.C. and Maia, M.D.A. (2014), "Ranking crowd knowledge to assist software development", in Roy, C.K., Begel, A. and Moonen, L. (Eds), *Proceedings of the 22nd International Conference on Program Comprehension, ACM, Hyderabad, 31 May-7 June*, pp. 72-82.

Tausczik, Y.R., Kittur, A. and Kraut, R.E. (2014), "Collaborative problem solving: a study of Mathoverflow", in Fussell, S., Lutters, W., Morris, M.R. and Reddy, M. (Eds), *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, Baltimore, MD, 15-19 February*, pp. 355-367.

Treude, C., Barzilay, O. and Storey, M.-A. (2011), "How do programmers ask and answer questions on the web? (NIER track)", *Proceedings of the 33rd International Conference on Software Engineering*, ACM, Waikiki, HI.

Treude, C., Figueira Filho, F., Cleary, B. and Storey, M.-A. (2012), "Programming in a socially networked world: the evolution of the social programmer", in Begel, A., Herbsleb, J.D. and Storey, M.-A. (Eds), *FutureCSD '12: Proceedings of the CSCW Workshop on the Future of Collaborative Software Development, Microsoft Research, Seattle, 12 February*.

Vasilescu, B., Serebrenik, A. and Van Den Brand, M.J. (2013), "The Babel of Software Development: Linguistic Diversity in Open Source", in Jatowt, A., Lim, E.-P., Ding, Y., Miura, A., Tezuka, T., Dias, G., Tanaka, K., Flanagin, A. and Dai, B. (Eds), *Social Informatics*, Springer International Publishing, pp. 391-404.

Vivaldi, J. and Rodríguez, H. (2010), "Finding domain terms using Wikipedia", *Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta*.

Wang, P. and Domeniconi, C. (2008), "Building semantic kernels for text classification using Wikipedia", in Li, Y., Liu, B. and Sarawagi, S. (Eds), *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Las Vegas, NV, 24-27 August*, pp. 713-721.

Wikipedia (2014a), "List of Wikipedias", Wikimedia Foundation Inc., available at: http://en.wikipedia.org/wiki/List_of_Wikipedias (accessed January 2016).

Wikipedia (2014b), "Wikipedia:size in volumes", Wikimedia Foundation Inc., available at: http://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes (accessed January 2016).

Wong, E., Jinqiu, Y. and Lin, T. (2013), "Autocomment: mining question and answer sites for automatic comment generation", *Automated Software Engineering (ASE), 2013 IEEE/ACM 28th International Conference*, *11-15 November*, pp. 562-567.

About the authors

Arash Joorabchi is currently a Postdoctoral Researcher in the Department of Electronic and Computer Engineering, University of Limerick, Ireland. He earned his 1st class Honours BSc in Computer Science from the Griffith College Dublin, Ireland in 2006 and received his PhD in 2010 from the Department of Electronic and Computer Engineering, University of Limerick, Ireland. The main focus of Arash's research is on developing and deploying text mining/analytics algorithms and techniques to automate the process of metadata generation in digital libraries and repositories. His major areas of interest include: data mining, knowledge organization, text analytics, digital libraries, and linked data. To find out more about Arash's work, visit www.csn.ul.ie/~arash/. Arash Joorabchi is the corresponding author and can be contacted at: arash.joorabchi@ul.ie

Michael English is a Lecturer in the Computer Science and Information Systems Department at the University of Limerick, Ireland. He received a BSc Hons in Mathematics and Statistics from the University College Cork in 1996. In 1999 he received an MSc in Computer Science from the University of Limerick and he received a PhD in Computer Science in 2007 from the same University. He has in excess of 15 years' experience in teaching computer science and computer programming at third level. His research interests are in the areas of computer science education, specifically in teaching and learning computer programming including identifying appropriate approaches to support at-risk students and software engineering, specifically software metrics, software quality and evolution and the analysis of source code. He has published his research in a number of peer-reviewed international journals and conferences.

Abdulhussain E. Mahdi is a Senior Lecturer at the Department of Electronic and Computer Engineering, University of Limerick (UL), Ireland. He is Director of the Regional Peer-Supported Learning Centre – UL and Joint Director of the ICT Learning Centre – UL. He is a Chartered Engineer (CEng), Member of the Institution of Engineering and Technology – UK (IET), and Member of the Engineering Council – UK. Dr Mahdi is a Graduate in Electrical Engineering from the University of Basrah (BSc 1st class Hons 1978) and earned his PhD in Electronic Engineering at the University of Wales – Bangor, UK in 1990. He is also a SEDA Accredited Teacher of Higher Education (University of Plymouth, UK 1998). His current research focuses on signal, speech, and NLP, data mining, machine learning and their applications in text analytics, information retrieval, management, and processing. His research interests also include peer-supported and collaborative learning, student-centred active learning, inquiry-based learning, and teaching and learning innovation and practices in ICT education. He has authored and co-authored more than 130 peer-reviewed journal articles, book chapters, and conference papers, and has edited one book. His published work has been cited more than 469 (Source: Google Scholar, September 2015) or 145 (Source: ISI-WoK, September 2015) times.