Web Forum Retrieval and Text Analytics: a Survey

Doris Hoogeveen  
University of Melbourne  
doris.hoogeveen@gmail.com

Li Wang  
Evernote, California  
li@liwang.info

Timothy Baldwin  
University of Melbourne  
tb@ldwin.net

Karin M. Verspoor  
University of Melbourne  
karin.verspoor@unimelb.edu.au
# Contents

1 Introduction 2  
1.1 Types of forums .................................................. 3  
1.2 A short history of forums ........................................ 7  
1.3 Scope and outline .................................................. 13  
1.4 Glossary ............................................................ 13  
1.5 Existing data sets .................................................... 16  

2 Post classification 18  
2.1 Post type classification ............................................ 19  
2.2 Question classification ............................................ 21  
2.3 Post quality assessment ............................................ 22  
2.3.1 Features for post quality classification ....................... 24  
2.3.2 Completeness and answerability ................................ 27  
2.4 Subjectivity and viewpoint classification ......................... 30  
2.4.1 Subjectivity and viewpoint classification in cQA archives ............................................................ 31  
2.4.2 Subjectivity and viewpoint classification in discussion forums ...................................................... 33  
2.5 Post classification summary ........................................ 34  

3 Post retrieval .......................................................... 35
3.1 Discussion forum post retrieval .......................... 36
3.2 CQA question retrieval ................................. 38
  3.2.1 Question retrieval: statistical translation models ... 40
  3.2.2 Question retrieval: topic models .................... 46
  3.2.3 Deep learning approaches to question retrieval ... 48
  3.2.4 Question retrieval: using category information ... 53
  3.2.5 Other question retrieval methods ................... 56
3.3 CQA answer retrieval .................................. 59
  3.3.1 Answer retrieval: topic models ..................... 60
  3.3.2 Answer retrieval: incorporating answer quality .... 61
  3.3.3 Answer retrieval: adding user information .......... 63
  3.3.4 Machine learning approaches to answer retrieval .. 64
  3.3.5 Other answer retrieval methods .................... 68
  3.3.6 Shared tasks on answer retrieval ................... 69
3.4 Post retrieval evaluation ................................ 71
3.5 Post retrieval summary .................................. 72

4 Thread level tasks ........................................... 73
  4.1 Task orientation and solvedness ....................... 73
  4.2 Thread discourse structure ............................ 75
    4.2.1 Thread linking structure recovery ................. 75
    4.2.2 Dialogue act tagging ................................ 76
    4.2.3 Thread partitioning ................................ 80
  4.3 Discussion forum thread retrieval .................... 81
  4.4 QA-pair extraction .................................... 85
  4.5 Thread summarisation .................................. 87
    4.5.1 Summarising discussion forum threads ............ 87
    4.5.2 Summarising cQA answers .......................... 92
  4.6 Thread level tasks summary ............................ 96

5 Social forum analysis .................................... 97
  5.1 User satisfaction ...................................... 97
  5.2 User and community analysis ......................... 100
  5.3 Expert finding ........................................ 103
    5.3.1 Question recommendation and question routing ... 106
  5.4 Social forum analysis summary ....................... 110
6 Conclusion
   6.1 Standardization and comparison of methods  111
   6.2 Challenges  112
   6.3 Open research questions  113

Acknowledgements  115

References  116
Abstract

This survey presents an overview of information retrieval, natural language processing and machine learning research that makes use of forum data, including both discussion forums and community question-answering (cQA) archives. The focus is on automated analysis, with the goal of gaining a better understanding of the data and its users.

We discuss the different strategies used for both retrieval tasks (post retrieval, question retrieval, and answer retrieval) and classification tasks (post type classification, question classification, post quality assessment, subjectivity, and viewpoint classification) at the post level, as well as at the thread level (thread retrieval, solvedness and task orientation, discourse structure recovery and dialogue act tagging, QA-pair extraction, and thread summarisation). We also review work on forum users, including user satisfaction, expert finding, question recommendation and routing, and community analysis.

The survey includes a brief history of forums, an overview of the different kinds of forums, a summary of publicly available datasets for forum research, and a short discussion on the evaluation of retrieval tasks using forum data.

The aim is to give a broad overview of the different kinds of forum research, a summary of the methods that have been applied, some insights into successful strategies, and potential areas for future research.
In this survey we will give an overview of a broad range of forum-related research. Forum research can be divided into two streams: discussion forums and community question-answering (cQA) archives. Both of these are websites that promote interaction and information sharing by the community, but they differ in their purpose, and because of that they often differ in their specific setup as well.

Forum data has been used for a large range of tasks and subtasks in information retrieval and natural language processing. Most of the tasks have to do with improving access to the rich information in the data, like post, question, or answer retrieval, thread summarisation, and expert finding. Subtasks cover specific aspects of the data and can be used to improve the results of the main tasks. Examples include dialogue act tagging, question and post type classification, post quality assessment, subjectivity and viewpoint classification, solvedness detection, thread type identification, topic detection, and user analysis. Forum research can also be used to improve the organization of the data, for instance by identifying duplicate questions, or categorizing posts.

In the remaining sections, we will present an overview of the different types of forums (§1.1), briefly discuss their history (§1.2), outline
1.1 Types of forums

In this section we will look at the differences between discussion forums and community question-answering archives. Both of these promote community interaction. Community question-answering archives are meant to help people to solve their problems and answer their questions. As soon as someone posts a good answer to a new question, the interaction is considered to be finished. Discussion forums on the other hand, are meant as a platform for people to discuss things.

This difference is not always strictly observed however. Some cQA archives contain questions like "Any1 from NY?", which do not express an information need, but rather a social need. Another example is requests for recommendations. Such questions do not have one correct answer and are therefore again more suited to discussion forums. Conversely, many factual questions and requests for help are posted on discussion forums, which might be more suitable for cQA archives.

Not much work has been published on the typology of forums. Choi et al. [2012] proposed a typology of online Q&A models consisting of four distinct types: community-based (e.g. Yahoo! Answers), collaborative (e.g. WikiAnswers), expert-based (e.g. the Internet Public Library (IPL) ‘Ask a Librarian’-service), and social (e.g. Twitter, which we do not consider to be a forum). Shah et al. [2014] placed the four cQA forum types in a hierarchical structure of Q&A services, which also includes face-to-face Q&A, and automatic Q&A services. Discussion forums are not present in either of these taxonomies. Several dimensions along which we can classify internet communication tools (including forums) are presented in Long and Baecker [1997]. While slightly outdated, it includes aspects like conversational style and audience membership, which are still valid today. Similar relevant dimensions or aspects can be found in Davies et al. [2005] (e.g. degree of interaction, motivation/orientation, size, maintenance, etc.).
In this survey we argue that forums exist on a spectrum with discussion threads on the one hand, where users have a high degree of freedom in what they post, and strict question-answering threads on the other, with heavy moderation to ensure only good answers are posted and threads are closed as soon as the question has been answered in a satisfactory way. In some cases the distinction is blurred. Linux Questions (http://www.linuxquestions.org/) for instance, looks like a forum, and has subforums dedicated to discussing Linux-related topics, but also focuses on answering questions. Yahoo! Answers (https://answers.yahoo.com/), a cQA archive, contains questions that look like they are intended to spark a conversation. An example can be found in Figure 1.1. This also illustrates the lack of moderation on Yahoo! Answers.

On the far end of the cQA side of the spectrum there are cQA sites with a high degree of moderation supplied by the community itself. On
1.1. Types of forums

such websites there is often a reward system in place for users that ask good questions and provide good answers. StackExchange is a good example of this. Figure 1.2 shows an example of a thread from the StackExchange Cooking site.

As can be seen in the example, a distinction is made between answers and comments. Comments are used to ask for clarification, correct people, offer small suggestions, or make general remarks or even jokes. Answers are reserved for genuine answers. The number of reputation points and other rewards the users have obtained is shown next to their name. In this way, active contributors and experts can be distinguished from new users. This can be one way for users to consider which answer is the best one. Users can also look at the number of up votes and down votes an answer has received. These votes are cast by the community to indicate the quality of answers (and questions).

Another characteristic of most cQA archives, and something that discussion forums do not offer, is that question askers are encouraged to choose one of the answers as the best answer. That way other users know that the information need has been satisfied and they can focus their efforts on other questions. Repeated questions can be linked to archived ones, and an active effort is made by the community to keep the answers focused and not to stray away from the question. When it does happen, the question is usually closed. This is very different from discussion forums, where some threads can ‘live’ for very long and no one is bothered by it. A classic example of this is the famous "i am lonely will anyone speak to me" thread posted in the Moviecodec.com branch discussion forum, The Lounge Forums, in 2004.\(^1\) It is still active today: more than twenty years since it was started.

Forums differ in how much access they offer to the outside world, but most of them make their content visible for everyone, while requiring people to sign up if they want to contribute. Some forums offer the option to sign up as an anonymous user. This makes the threshold to

\(^1\)https://www.loungeforums.com/on-topic/i-am-lonely-will-anyone-speak-to-me-2420/. It is more than 2000 pages long. Several magazines and newspapers have featured this thread. See for more information https://en.wikipedia.org/wiki/I_am_lonely_will_anyone_speak_to_me.
contribute lower. In some forums that is seen as a good thing, because it lowers the bar of entry, but in forums that want to create a steady community of people that contribute regularly, these kinds of one-off contributions are discouraged. Having a system where people need to sign up before they can participate has the added benefit of making it difficult for bots to post spam, and it allows for personalisation of the forum. Some forums even offer member pages with all kinds of meta data such as when they became a member, how active they are, reputation points, question and answer history, and all the subforums they participate in, or topics they have expertise in. StackExchange\textsuperscript{2} is once again a good example of this.

While many discussion forums explicitly show the discourse structure of the thread, i.e., which post is a reply to which earlier post, (see Figure 1.3 for an example), this is not always the case (see Figure 1.4). Quoted posts, allowed by some forums and illustrated in Figure 1.5, can be used to retrieve at least part of the discourse structure. We discuss this in §4.2.

CQA archives only have a simple two-part discourse structure, between a question and each of its answers. The original order of the answers is often not preserved. Instead, they are usually ordered based on the number of votes they have received from the community, with the answer that was accepted as the correct one by the question asker at the top.

1.2 A short history of forums

One of the earliest examples of a community question-answering service is The Straight Dope\textsuperscript{3} founded in 1973. It started out as a column in several American newspapers, but these days it also has an online forum where people can ask questions and receive answers. The setup is closer to a discussion forum than a CQA archive however, with several subforums specifically created for discussion, such as the Elections subforum.

\textsuperscript{2}http://stackexchange.com/
\textsuperscript{3}http://www.straightdope.com/
Introduction

Figure 1.3: An example of a discussion forum thread with explicit discourse structure. Source: Reddit, https://www.reddit.com/r/Showerthoughts/comments/5403tk/the_reason_why_tomato_soup_and_grilled_cheese_is/, accessed on 24th of September 2016.
1.2. A short history of forums

Another early example is the Internet Oracle\(^4\) founded in 1989. It specialises in humorous answers. Although it is a community question-answering service, questions and answers are submitted and distributed via e-mail.

Discussion forums also started to appear in the late 1980s. The Delphi Forums\(^5\) was created in 1983 and is one of the earliest forums; it is still active today. Online discussion forums have their origins in bulletin boards and newsgroups such as Usenet, which has been around since 1980.

\(^4\)http://internetoracle.org/
\(^5\)http://www.delphiforums.com/
1.2. A short history of forums

In the 1990s several cQA archives emerged. For instance:

- The Madsci Network:\(^6\) It is heavily moderated and questions are all answered by scientists, rather than being open to anyone willing to contribute.

- Experts-Exchange\(^7\): This site is specifically for technology experts. It started out as purely community question-answering, but has expanded and now also offers help with code reviews, connecting freelancers to jobs, educating people, and live chat with an expert.

- 3form:\(^8\) focuses on finding solutions to problems, rather than answers to questions. That is, questions are requests for information, either factual or not, while problems are questions for help in solving a particular issue.

Discussion forums also grew in popularity. In 1994 the W3C introduced WWW Interactive Talk (WIT),\(^9\) a discussion forum that followed a set of design principles to display online discussions in such a way that it was easy to see which different topics were being discussed, and which points had been settled or not. Before WIT, many discussion forums suffered from the problem of people posting the same arguments over and over again, because there was no clear overview of a full thread. Although this was a step forward, and many alternatives sprang from this, to a certain extent we are still struggling with similar issues today.

In the first decade of the 2000s a large number of new cQA archives appeared, many of which are still extremely popular today: Baidu Knows,\(^10\) WikiAnswers/Answers.com,\(^11\) Quora,\(^12\) Naver Knowl-

\(^6\)http://www.madsci.org/ started in 1995 and still going.
\(^7\)https://www.experts-exchange.com/ started in 1996 and still going.
\(^8\)http://3form.org/ started in 1998 and still going.
\(^9\)Official website: https://www.w3.org/WIT/, and more information can be found at http://speed.eik.bme.hu/help/html/Special_Edition-Using.cgi/ch17.htm#WWInteractiveTalk.
\(^12\)https://www.quora.com/ started in 2009.
edge Search,\textsuperscript{13} Yahoo! Answers,\textsuperscript{14} and the StackExchange\textsuperscript{15} website, especially StackOverflow.\textsuperscript{16} The only notable exception is Google Answers\textsuperscript{17} which was started in 2002 but discontinued in 2006. Many of these large cQA archives are in English, but not all of them: Naver is Korean, and Baidu Knows is Chinese.

One specific example of a space where forums have been used and found to be helpful is education. There are several online cQA archives dedicated to questions about topics taught in schools. An example of this is Brainly,\textsuperscript{18} which has the slogan “For students. By students.” The idea is that students help each other to learn. Other examples are Chegg,\textsuperscript{19} and Piazza.\textsuperscript{20} Lang-8\textsuperscript{21} is a language learning platform that has many similarities to forums. Users write posts in a language they are learning. Native speakers of that language will then correct the post sentence by sentence and comment on it. The original poster can reply to the corrections, and other native speakers can join in the conversation too, to discuss linguistic constructs or explain semantic or syntactic points.

Many learning management systems include a forum to enable students to start discussions online, or ask questions. This is considered to be a vital ingredient of MOOCs for instance, where the number of students is so large that it is not possible for them to individually get in touch with the professor or tutors, and forums offer an alternative to ask for help or discuss the subject matter. In such a setting, the forums are used both as a cQA platform and as a discussion forum. One MOOC platform, EdX,\textsuperscript{22} has recognised this dual nature of MOOC forums and allows people to choose what kind of post they make: a

\begin{footnotesize}
\begin{enumerate}
\item http://kin.naver.com/index.nhn started in 2002.
\item https://answers.yahoo.com/ started in 2005 and formally known as Yahoo! Q&A.
\item http://stackexchange.com/ started in 2008.
\item https://stackoverflow.com/, the first cQA site of the StackExchange network.
\item http://answers.google.com/answers/. It grew out of Google Questions and Answers which was started in 2001.
\item http://brainly.com/
\item https://www.chegg.com/
\item https://piazza.com/
\item http://lang-8.com/
\item https://www.edx.org/
\end{enumerate}
\end{footnotesize}
1.3. Scope and outline

In this survey we will describe research into automated analysis of forum data. That includes data from both discussion forums (also called web user forums; see, for instance, [Wang et al., 2013b]) and community question-answering (cQA) archives. These two forum types share a number of characteristics (as discussed in §1.1), which are not shared with other (semi) threaded discourses, like chat discussions, email threads, product reviews, or frequently asked question (FAQ) pages. These are therefore outside the scope of this survey.

At the start of this section we mentioned several tasks and subtasks. Each of these will be discussed in the following sections, divided into post classification (§2), post retrieval (§3), thread level tasks (§4), and social forum analysis or user studies (§5).

Previously published survey articles include Shah et al. [2009], who present an overview of early research in the cQA field, Gazan [2011], Li [2014], and Srba and Bielikova [2016], who all present an overview of cQA related research. Srba and Bielikova [2016] is the most recent and most comprehensive survey, discussing 265 research papers published before February 2015. They also show that the number of publications in this field has increased each year.

This survey covers 450 papers published until November 2016, and distinguishes itself from earlier survey papers by including discussion forums, instead of focusing on cQA archives only.

1.4 Glossary

The same or similar concepts sometimes appear in the literature under different names. We will try to use the same terminology for each con-

---

23 http://blog.edx.org/navigating-two-kinds-online-discussion
cept throughout this survey. This section summarises the important terminology we will use.

**Thread:** we use the term “thread” to refer to forum discussion threads, or a question on a cQA forum together with all of its answers (and comments). In discussion forums this is the full thread, which may span multiple pages (see **Page** below).

**Page:** in discussion forums, threads can sometimes become very large. If this happens, instead of displaying the full thread, only a certain number of posts are displayed at a time. So threads are divided into smaller units for easier display. Such chunks are called “pages”.

**Post/message:** the terms “post” and “message” are often used interchangeably in the research community to refer to each posting in a forum thread. In this survey we use “post” to denote forum thread post. The term “post” can also be used to refer to either the question post in a cQA archive, or an answer post. We use it as a general term when we want to refer to any text posted by a user, regardless of whether it is an initial post or question post, or an answer post. In situations where it matters we will distinguish clearly between the two, by calling them “initial post” (or “question post”) and “answer post”.

**Initial post:** this refers to the first post in a discussion forum thread, which starts a discussion. In the literature, it is sometimes also called the “root post/message” or “first post/message”.

**Question post:** this refers to the first post in a cQA thread, in which a question is asked. All other posts in a cQA thread are answers to this post.

**Answer post:** this refers to any post in a cQA thread that is not the question post, but rather a response to a question post.

**Word/term:** in this survey, “word” and “term” are used interchangeably to indicate a word unit in a post.
1.4. Glossary

**Thread initiator:** the user who starts a new discussion thread (in discussion forums), or who posts a question (in cQA archives). This is the person that writes the Initial post or Question post. In a cQA context we will sometimes refer to this person as the “question asker”.

**Quoted text:** in discussion forums a user may sometimes quote content from previous posts or email messages in his/her post. This quoted content is called “quoted text”. In cQA archives, quoted material often comes from other threads or from technical documentation. An example from a discussion forum can be found in Figure 1.5.

**Comment:** in some cQA archives, users can write comments to posts, in addition to answers. These two kinds of posts (comments and answers) serve a slightly different purpose. Answers are supposed to directly answer the question, while comments can be used to correct someone, ask for clarification on a certain point, make a small addition to a post, or provide similar short contributions that are not standalone answers.

**Thread structure:** The structure of a discussion forum thread can be viewed as a tree, with the initial post at the top, and reply posts branching out below it. Each post is placed below the post it responds to. This structure can be explicit, like in Figure 1.3, or not, like in Figure 1.4.

As background information we would like to very briefly introduce some IR evaluation metrics here, which will be mentioned in different places throughout this survey. Many different evaluation metrics are used for IR tasks using forum data, i.e. post retrieval, and IR in general. For instance, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) [Voorhees, 1999], Precision@n, nDCG [Järvelin and Kekäläinen, 2002], AUC (precision–recall or ROC), and Rank-Biased Precision [Moffat and Zobel, 2008]. Of these, MAP is the most widely used. It is the mean of the average precision at a given
cut-off point, calculated over all the queries in a set. The average precision is shown in Equation 1.1, in which \( N \) is the cut-off point, \( P \) is the precision, and \( R \) is an indicator of whether the document retrieved at \( i \) is relevant or not.

\[
\text{AP@N} = \frac{\sum_{i=1}^{N} P(i) \cdot R(i)}{\text{# of relevant documents}}
\]  

(1.1)

1.5 Existing data sets

The field of forum related research has long suffered from a lack of publicly available datasets, but this is slowly changing. Over the years, many researchers have constructed their own sets using web forum crawling techniques, for instance using methods described in Wang et al. [2008] or Yang et al. [2009a]. Recently, some forums have started making (part of) their data available to the research community, and many top-tier conferences (e.g. the AAAI International Conference on Web and Social Media) encourage their authors to share their data and provide data sharing services specifically for this purpose. An overview of a large number of public and private datasets used in forum research can be found in Hoogeveen et al. [2015]. In this section we will present only the most important ones, which are openly available for research purposes. They are summarised in Table 1.1.
1.5. Existing data sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The WikiAnswers Corpus</strong></td>
<td>30M clusters of questions from WikiAnswers,(^{24}) tagged as paraphrases by users. Around 11% of them have an answer.</td>
</tr>
<tr>
<td><strong>TREC 2015 LiveQA data</strong></td>
<td>1000 Yahoo! Answers questions used as queries in the TREC 2015 LiveQA task, including answer strings from systems, with human judgements.</td>
</tr>
<tr>
<td><strong>The SemEval Task 3 cQA Dataset</strong></td>
<td>2900 English questions and answers from the Qatar Living Forum,(^{25}) and 1500 Arabic ones from the Fatwa forum on IslamWeb.(^{26})</td>
</tr>
<tr>
<td><strong>StackExchange dump</strong></td>
<td>A periodical dump of all the data on StackExchange, in XML format.</td>
</tr>
<tr>
<td><strong>CQADupStack</strong></td>
<td>All the data of twelve StackExchange forums, in JSON format.</td>
</tr>
<tr>
<td><strong>MSR Challenge Dataset</strong></td>
<td>Stripped version of a StackOverflow dump, in XML and postgresql formats.</td>
</tr>
<tr>
<td><strong>The NTCIR-8 cQA dataset</strong></td>
<td>1500 questions and answers from Yahoo! Chiebukuro, the Japanese version of Yahoo! Answers, between April 2004 and October 2005.</td>
</tr>
<tr>
<td><strong>The Reddit Comment Corpus</strong></td>
<td>A periodical dump of all the comments. Some of it contains sentiment annotations.</td>
</tr>
<tr>
<td></td>
<td><a href="https://www.reddit.com/r/datasets/comments/590re2/updated_redic_comments_and_posts_updated_on/">https://www.reddit.com/r/datasets/comments/590re2/updated_redic_comments_and_posts_updated_on/</a></td>
</tr>
<tr>
<td><strong>The Quora Dataset</strong></td>
<td>400,000 question pairs, annotated for duplicates. Released on 25/01/2017.</td>
</tr>
<tr>
<td></td>
<td><a href="https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs">https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs</a></td>
</tr>
</tbody>
</table>

Table 1.1: An overview of publicly available forum data sets.
References


References


References


References


References


References


References


Li Gao, Yao Lu, Qin Zhang, Hong Yang, and Yue Hu. Query Expansion for Exploratory Search with Subtopic Discovery in Community Question Answering. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 4715–4720. IEEE, 2016.


Iryna Gurevych, Delphine Bernhard, Kateryna Ignatova, and Cigdem Toprak. Educational Question Answering Based on Social Media Content. In *Pro-
cedings of the International Conference on Artificial Intelligence in Edu-


References


References


References


References


References


References


References

