JIS

Article

Automatic Mapping of User Tags to Wikipedia Concepts: the Case of a Q&A Website - StackOverflow

Journal of Information Science
1–15

DOI: 10.1177/0165551510000000
jis.sagepub.com
SAGE

Arash Joorabchi
Department of Electronic and Computer Engineering, University of Limerick, Ireland

Michael English
Department of Computer Science and Information Systems, University of Limerick, Ireland

Abdulhussain E. Mahdi
Department of Electronic and Computer Engineering, University of Limerick, Ireland

Abstract
The uncontrolled nature of user-assigned tags makes them prone to various inconsistencies caused by spelling variations, synonyms, acronyms, and hyponyms. These inconsistencies in turn lead to some of the common problems associated with the use of folksonomies such as the tag explosion phenomenon. Mapping user tags to their corresponding Wikipedia articles, as well-formed concepts, offers multi-facet benefits to the process of subject metadata generation and management in a wide range of online environments. These include normalization of inconsistencies, elimination of personal tags, and improvement of the interchangeability of existing subject metadata. In this article, we propose a machine learning-based method capable of automatic mapping of user tags to their equivalent Wikipedia concepts. We have demonstrated the application of the proposed method and evaluated its performance using the currently most popular computer programming Q&A website, StackOverflow.com, as our test platform. Currently, around 20 million posts in StackOverflow are annotated with about 37,000 unique user tags, from which we have chosen a subset of 1,256 tags to evaluate the accuracy performance of our proposed mapping method. We have evaluated the performance of our method using the standard information retrieval measures of precision, recall, and F_1. Depending on the machine learning-based classification algorithm used as part of the mapping process, F_1 scores as high as 99.6% were achieved.

Keywords
Semantic mapping; subject metadata; user tags; StackOverflow; Wikipedia

1. Introduction

In recent years, user tagging (a.k.a. crowdsourced tagging, social tagging, collaborative tagging) has become a popular approach to generate subject metadata for a wide range of online materials such as, photos (e.g., Flickr[1]) videos (e.g., Vimeo[2]), books (LibraryThing[3]), and research papers (e.g., CiteULike[4]). This may be attributed to the explosive growth of online content which makes the task of authoritative classification and subject indexing, traditionally carried out by professional cataloguers in library settings, infeasible in many cases. As an alternative to the professional indexing with controlled vocabularies, user tagging relies on user communities to collaboratively index resources of their interest with uncontrolled vocabularies (a.k.a folksonomies) [1]. The crowdsourced nature of user tagging reduces the cost of indexing significantly and makes it a viable option for subject metadata generation in many online settings. However, compared to controlled vocabularies such as library classification systems and subject headings, which are developed and maintained by experts, folksonomies suffer from various inconsistency issues and offer a lower indexing quality [2]. The uncontrolled nature of folksonomies makes them prone to inconsistencies caused by spelling variations, synonyms, acronyms, and hyponyms. These inconsistencies in turn could lead to problems such as "tag explosion", where a small subset of tags from a folksonomy is used to annotate a great majority of items in a collection and the remaining tags are

Corresponding author:
Abdulhussain E. Mahdi, Department of Electronic and Computer Engineering, University of Limerick, Limerick, Republic of Ireland.
Email: Hussain.Mahdi@ul.ie

used only minimally [3]. Another well-known issue with user-assigned tags is the substantial usage of personal tags which have no subject indexing value for the community, e.g., "to read", "unread" [4]. Investigating the substitutability of social tagging with professional indexing, Lee and Schleyer [5] compared user tags with Medical Subject Headings (MeSH)[5] terms for a set of biomedical papers that appeared in both CiteULike and MEDLINE[6]. The results of this study showed that social tags and MeSH terms have little overlap and embody largely heterogeneous understanding of items; hence the authors concluded that social tagging is no substitute for controlled indexing. In a similar work, Wu et al. [6] studied the relationship between social tagging and controlled vocabulary-based indexing in the domain of information science for both English and Chinese languages. On the basis of the results of their study, the authors concluded:

> Overall, despite the limitations of applying social tagging in cataloguing and indexing, we do believe that it has the potential to become a complementary source to expand and enrich controlled vocabulary systems. With the help of future technology to check consistency and promote features related to controlled vocabulary in social tags, a hybrid cataloguing and indexing system that integrates social tags with controlled vocabulary would greatly improve people's abilities to organize and access information resources.

In the current situation, where professionally assigned subject metadata is rarely available due to its high cost and user generated subject metadata is abundant but low in quality, association and mapping of user tags to their semantically corresponding terms and subject headings in controlled vocabularies could significantly improve the quality of existing subject metadata. Furthermore, implementing such integrations would improve the interchangeability and augmentability of existing subject metadata across platforms, and pave the way for the design and development of semantically enhanced information retrieval systems for clustering, ranking, and recommendation of items and records [7, 8]. As Noruzi [9] emphasizes,

> A controlled vocabulary for a folksonomy-based system is essential to ensure tagging consistency across the database and between taggers. This may be a thesaurus or subject headings. By controlling the vocabulary using a thesaurus, tags are standardized and related resources are collocated for ease of discovery by the end-user.

In this context, Wikipedia may serve as an effective target controlled vocabulary for mapping user tags to. Wikipedia is the world's largest free online encyclopedia. The English Wikipedia alone currently contains more than four million articles [10]. 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, collectively called Wikipedians. Despite the 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 U.S. 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[7] and Google Books[8] with a large margin [11]. Our justification for adopting Wikipedia as a controlled vocabulary is described and argued in Section 2.

Based on above, the objective of this work is to design and develop a robust automatic method for mapping user tags to Wikipedia concepts. In specific, we propose a machine learning-based method for mapping user tags from the most popular Q&A website in the field of computer programming, StackOverflow[9], to their corresponding concepts in Wikipedia. The problem of discovering and associating semantics to user tags has been tackled in other domains before. For example, Yi and Chan [12] investigated the linking of user tags in a popular collaborative tagging system, Delicious[10], to the Library of Congress Subject Headings (LCSH)[11] using a word-matching-based method. Golub et al. [13] investigated enhancing user tags with automated keywords from the Dewey Decimal Classification (DDC)[12] system. Angeletou et al. [14] proposed an automatic approach to associate user tags in Flickr with senses from WordNet[13]; see [15] for a review of similar work. However, to the best of our knowledge, this is the first attempt to automatically map user tags in one of the websites from the StackExchange[14] network to their corresponding Wikipedia concepts. The content of these Q&A websites and their user tags present a set of unique challenges and opportunities to address, and the new mapping method proposed in this work aims to target the requirements of this particular domain.

The rest of the article is organized as follows: Section 2 discusses the application of Wikipedia as a controlled vocabulary and compares it with expert-built controlled vocabularies in the context of subject metadata generation. Section 3 introduces the StackExchange network and its most popular website, StackOverflow. It describes the current subject metadata generation method used in these platforms, and provides some statistics in relation to the quantity and quality of their existing user tags. Section 4 describes our automatic mapping method and its implementation details. Section 5 describes the evaluation process and presents its results. This is followed by Section 6 which provides a conclusion along with a summary account of planned future work.

2. Wikipedia as a Controlled Vocabulary

In recent years, Wikipedia has received a lot of attention from researchers working in the field of information retrieval and knowledge management [16]. 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 [17], text classification [18], text clustering [19], event detection [20], topic indexing [21], and semantic relatedness measurement [22]. As a controlled vocabulary, Wikipedia offers a number of advantages over traditional controlled vocabularies:

Extensive coverage and comprehensiveness: the English Wikipedia currently contains over 4 million articles covering subjects in all aspects of human knowledge and growing. This allows adapting Wikipedia as a general thesaurus on its own or in conjunction with traditional expert-built thesauri such as WordNet [23]. Furthermore, the substantial coverage of Wikipedia in various knowledge domains has enabled researchers to derive high quality domain-specific thesauri from it [24, 25].

Up-to-date: 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 [26]. As a controlled vocabulary, Wikipedia is able to keep pace with swiftly changing domains via continuous addition of new concepts. Whereas, traditional expert-built thesauri, which are constructed by professional indexers and taxonomists, go through update cycles at much lower pace. For example, a new version of Agrovoc[15], which is an expert-built thesaurus in the domain of agriculture, is released every two years[16]. This issue becomes more prominent in case of fast moving domains and emerging sciences, such as information and computer science, which undergo rapid evolution.

Rich description: Wikipedia articles provide rich descriptive content for the represented concepts. Whereas, traditional controlled vocabularies offer little or no description for their terms and subject headings. For example, the LCSH authority record for the subject heading "Metadata"[17] offers only two pieces of information about this subject: (a) the subject may also be referred to by "data about data" and "meta-data"; and (b) the subject is related to two more specific subjects of "Dublin Core" and "Preservation metadata". In contrast, the Wikipedia article for the concept of "metadata"[18] provides a rich description for the concept including its definition, variations, and applications, complemented with links and references to relevant materials and related concepts. This feature of Wikipedia enables users to find descriptive information about unfamiliar subject indexes which they may encounter while browsing and exploring a collection. Furthermore, Wikipedia allows users to examine the history of each concept and review the discussions and debates around its definition and development.

Rich semantics: according to part 1 of the international standard for thesauri (ISO 25964-1)[19], a compliant thesaurus should capture and encode three main types of relationship between concepts: (a) equivalence relations between synonyms and near-synonyms, e.g. car and automobile, (b) hierarchical relations between broader and narrower concepts, e.g. vehicle and car, (c) 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: (a) 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), (b) Similar to the notion of "Related Terms" in traditional controlled vocabularies, related articles in Wikipedia are connected via hyperlinks (i.e., associative relations), (c) 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). Wikipedia addresses the problem of word-sense ambiguity using disambiguation pages which list all possible senses of an ambiguous term and provide links to the corresponding concepts for each unique sense, e.g. Java (programming language), Java (town), Java (band). This is equivalent to "scope notes" in traditional controlled vocabularies which are used to clarify the boundaries of a concept and its intended use for indexers and searchers.

Multilingual: as of July 2014, Wikipedia exists in more than 287 languages. Wikipedia has more than one 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 [27].This high level of multilingualism in Wikipedia allows researchers and practitioners to adapt it as a multilingual controlled vocabulary in various information retrieval and indexing applications. For example, Melo and Weikum [28] integrated all editions of Wikipedia and WordNet into a

single coherent taxonomic class hierarchy called MENTA (Multilingual Entity Taxonomy) that describes 5.4 million entities.

The main purpose of the Wikipedia project is to build a free online encyclopaedia and, as such, it has some limitations compared to traditional controlled vocabularies:

- Articles in Wikipedia are connected to each other via an extensive network of hyperlinks which can be mined for discovering associative relations between the represented concepts. However, these links do not always explicitly equate to the notion of "Related Terms" in traditional controlled vocabularies. Also, the types of links (e.g. part-of, member-of, instance-of) are not encoded in Wikipedia.
- Traditionally expert-built classification systems and taxonomies, such as the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) adhere to a hierarchical tree structure. However, Wikipedia's classification system has a loose semi-hierarchical directed-graph structure which allows articles to belong to multiple categories, and categories to have multiple parents. Also, the collaborative and crowdsourced nature of taxonomy development and categorization work in Wikipedia makes it prone to some level of noise. For example, our analysis of the Wikipedia dump used in this study and those done by others (e.g., see [29]) have shown the existence of self-loops (C1➔C1), direct-loops (C1➔C2➔C1), and indirect-loops (e.g., C1➔C2➔C3➔C1) among some categories in the Wikipedia's classification graph. Consequently, hierarchical relations (BT/NT) among Wikipedia concepts are subject to some level of noise.

The accuracy of Wikipedia articles has always been subject of controversy due to Wikipedia's open editing policy [30]. However, an investigation conducted by Nature in 2005 [31] 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 [32]. Irrespective of these controversies, the poor editorial quality of some Wikipedia articles and their occasional factual inaccuracies do not seem to have a significant adverse effect on the quality of the controlled vocabularies derived from Wikipedia. For example, Milne et al. [24] 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 [25] 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. Xu et al. [33] investigated the application of Wikipedia thesaurus knowledge to improve the performance of contextual web advertising, and showed that their approach can substantially improve the performance of ad selection and outperform the conventional contextual advertising matching approaches. Macías-Galindo et al. [34] described a process for constructing domain-specific ontologies using concepts and associations imported from WordNet and Wikipedia. They evaluated the constructed ontologies by asking human subjects to rate the domain-relevance of the concepts included in each ontology on a 3-point scale. They reported achieving precision values between 71% and 88% and recall values between 37% and 95%. These and a substantial 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 [35-37].

3. Stack Exchange & Stack Overflow

StackExchange is a network of Q&A websites, each covering a specific topic (e.g., mathematics, physics, biology) in broad areas, such as technology, science, and business. According to Alexa[20], it currently ranks at number 170 in terms of global traffic. The network currently contains 119 "topic" websites and the same number of corresponding "meta" websites. Each topic website covering a specific subject has an accompanying meta website dedicated to discussions regarding its management and maintenance. 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 (e.g., allocation of rewards in the form of badges and reputation scores) to encourage and stimulate community participation [38].

Created in 2008, StackOverflow was the first website in the StackExchange network, 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. StackOverflow currently has over 3.5 million registered users. Since its inception, more than 8 million questions have been posted on the site and over 14 million answers have been provided [39], all contributing to a large knowledge repository of computer programming and software development. Parnin and Treude [40] investigated the documentation resources that programmers use by analysing Google search results for a popular API (jQuery) and found that StackOverflow appears (at least once) on the first results pages of 84% of the tested search queries. Although one might argue that this evidence only proves the popular usage of StackOverflow for discussions on a particular technology, Barua et al. [41] show that StackOverflow covers and has active discussions on a wide range of technologies. Currently an average of 7000 questions are posted on the site daily, and as of August 2010, StackOverflow had an answer rate above 90% and a median answer time of only 11 minutes [42]. According to Nasehi et al. [43], as of February 2012, the median time of accepted answers being posted on the site was 24 minutes and in the first hour 70% of questions received their first answer. Before the advent of social Q&A websites, online forums and threaded discussions were the most commonly used mechanism for Q&A. The main problem with this approach is that useful information is mixed with redundant and irrelevant information. Social Q&A websites such as StackOverflow, on the other hand, make use of collaborative filtering to rank best answers and show them up front, saving users time and effort [44].

When posting a new question on StackOverflow, a user is asked to provide 1 to 5 tags for the question. Each tag has a dedicated wiki page in which users provide additional information about the tag itself and its appropriate use cases. The wiki page contains two main information elements: a short excerpt and a full description. Excerpts are the most visible part of tag wikis. They are shown when hovering over a tag, on the tags page, and in the auto complete suggestions when adding tags to a question. The full description part of a tag wiki complements its excerpt by providing more background and detailed information about the tag. Figure 1 shows the wiki page of a sample tag, rapidminer[21], including its excerpt (appearing inside a grey rectangle), full description, and some additional information such as usage frequency, creation date, number of views, etc.

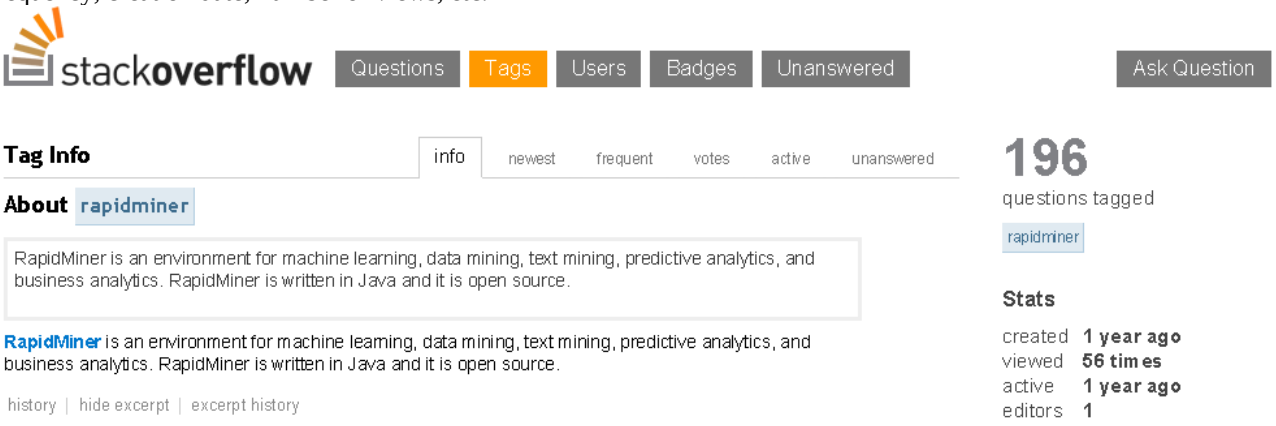

Figure 1. A sample tag's wiki page.

The StackExchange network, which the StackOverflow website is part of, has adopted an open data policy and publishes an anonymized dump of all its user-contributed content periodically on archive.org[22]. Each site is formatted as a separate archive consisting of zipped XML files for Posts, Users, Tags, Votes, Comments, Badges, PostHistory, and PostLinks (for complete schema information, see the readme file[23]). In this study we have used the latest release of StackOverflow data dump published in May 2014. In order to easily interact with this data, we uploaded the XML files of interest for our purposes in this work (Posts.xml, Tags.xml) into an open-source native XML database engine, eXist-db [45]. Considering the large size of the data (28GB), this enables us to easily and efficiently iterate, search, and retrieve the StackOverflow content by submitting rich XPath and XQuery queries to the database. Analysing this content we found a total of 7,214,697 questions, 12,609,623 answers, and 36,942 unique tags out of which 23,702 had corresponding wiki pages. Such large number of user tags in StackOverflow is a testimony to our argument in the previous sections regarding issues arising from the uncontrolled nature of user tags and folksonomies. Among these issues is the problem of "tag explosion", where a small subset of tags is used to annotate a great majority of items in the collection and the remaining tags are only used minimally. We examined the usage frequency of tags in StackOverflow

to verify if the problem of tag explosion exists in this case. Table 1 summarizes our findings in relation to the usage frequency of user tags within StackOverflow.

Table 1. Usage frequency of user tags in StackOverflow.

Usage Frequency Range	Number of Tags in the Range	Number of Tags in the Range (%)	Cumulative Total Usage	Cumulative Total Usage (%)
[0, 10]	10,311	27.91%	49,243	0.23%
(10, 100]	16,463	44.56%	628,054	2.95%
(100, 1000]	7,921	21.44%	2,476,832	11.63%
(1000, 10000]	1,969	5.33%	5,539,839	26.02%
(10000, 100000]	258	0.70%	6,223,886	29.23%
(100000, 635338]	20	0.05%	6,376,494	29.94%
[0, 635338]	36,942	100.00%	21,294,348	100.00%

We found the tag "C#" with a total usage frequency of 635,338 to be the most frequently used tag in StackOverflow. The data presented in Table 1 clearly exhibit a case of tag explosion, where 6.08% of tags (i.e., those with a usage frequency between 1000 and 635338) have been used to annotate 85.19% of questions, and the remaining 93.92% of tags (i.e., those with a usage frequency between 0 and 1000) have been used to annotate only 14.81% of questions. Identifying all the reasons behind this phenomenon in StackOverflow and similar settings, where user tagging leads to tag explosion, is beyond the scope of this work. However, our preliminary study of user tags in StackOverflow shows that at least part of the problem is caused by the over usage of version specific tags. For example, in the case of questions related to NetBeans, which is an integrated development environment for developing primarily with Java, in addition to the primary tag of "NetBeans", 12 version specific tags have been created and used, e.g., netbeans-7.0, netbeans-7.1, netbeans-7.2. As we found out, in many cases, the posted questions did not warrant usage of version specific tags and more importantly were missing the generic "NetBeans" tag, which hinders the search and retrieval efficiency. We believe this issue along with some of the other common problems stemmed from the uncontrolled nature of folksonomies (discussed in Section 1) could be addressed by mapping user tags to Wikipedia and its well-defined concepts as a controlled vocabulary.

4. Methodology

Our approach to automatic mapping of user tags in StackOverflow to their corresponding Wikipedia concepts comprises of two main stages: (a) identifying all the Wikipedia concepts appearing in the wiki page of the tag to be mapped; and (b) binary classification of detected concepts into equivalent or non-equivalent concepts.

In the first stage, we utilize an open-source toolkit called Wikipedia-Miner [46] for detecting Wikipedia concepts occurring in the tags' wiki pages. Wikipedia-Miner effectively unlocks Wikipedia as a general-purpose knowledge source for Natural Language Processing (NLP) applications by providing rich semantic information on concepts and their lexical representations. We use the topic detection functionality of the Wikipedia-Miner to identify all the Wikipedia concepts (i.e., Wikipedia articles) whose descriptor or non-descriptor lexical representations occur in a tag's wiki page. For example, the 16 Wikipedia concepts detected in the wiki page of the sample tag, rapidminer, shown in Figure 1 are: RapidMiner, Environment (biophysical), Machine, Machine learning, Learning, Data, Data mining, Mining, Text mining, Predictive analytics, Prediction, Analytics, Business analytics, Business, Java (programming language), Open source.

The result of concept detection stage is a set of candidate Wikipedia concepts from which only one could be the true match for the tag. The aim of the second stage is to find this true match using a Machine Learning (ML) based binary classifier which classifies each candidate concept as either "equivalent" or "non-equivalent" to the tag. In practice, each tag has only one true matching Wikipedia concept, and therefore only one of the detected concepts in its wiki page could truly belong to the equivalent category and the rest should be classified into the non-equivalent category.

Therefore, in case of a successful mapping, the number of True Positive cases (i.e., correctly identified) is 1, and the number of True Negative cases (i.e., correctly rejected) is *n*-1, where *n* represents the number of detected candidate concepts. For example, in case of the sample tag rapidminer, if mapped correctly to the RapidMiner[24] concept in Wikipedia, then *TP*=1, *TN*=16-1.

To build the above proposed classifier we need to (a) define a set of appropriate features for Wikipedia concepts, and (b) manually map a collection of sample tags to their equivalent Wikipedia concepts for training a classification model and testing its prediction performance.

4.1. Features for Wikipedia Concepts

In order for an ML-based classifier to identify a tag's equivalent Wikipedia concept among all the concepts detected in the tag's wiki page, a set of features for capturing the distinguishing properties of the concepts belonging to the equivalent category is required. We have devised a set of nine positional, statistical, and semantic features to capture and reflect various characteristics of those candidates which have the highest probability of belonging to the equivalent category:

(1) Frequency: the occurrence frequency of the candidate concept (i.e., descriptor of the concept) and its synonyms and alternative lexical forms/near-synonyms (i.e., non-descriptors of the concept) in the tag's wiki page. The Frequency values are normalized by dividing them by the highest Frequency value in the wiki page. We expect the tag's equivalent concept to have a relatively higher occurrence frequency compared to other candidate concepts identified in the tag's wiki page. The effectiveness of this feature is well proven in similar information retrieval and text mining applications such as automatic topic indexing [47] and keyword extraction [48].

(2) First Occurrence: the distance between the start of the tag's wiki page and the first occurrence of the candidate concept, measured in terms of the number of characters and normalized by the length of the page. This feature reflects the observation that the tag's equivalent concept tends to appear for the first time in the first line/paragraph of the wiki page. The effectiveness of this feature is proven in similar applications such as keyphrase extraction [49, 50], where candidates occurring close to the beginning of a document are shown to have a higher keyphraseness probability.

(3) Last Occurrence: the distance between the end of the tag's wiki page and the last occurrence of the candidate concept, measured in terms of the number of characters and normalized by the length of the page. This feature reflects the observation that in a considerable number of cases the tag's equivalent concept may reappear at the end of the tag's wiki page in form of hyperlinks to external information sources. This feature is proven to be effective in similar applications, where the candidate concepts occurring close to the end of a document, e.g., conclusion and reference sections, are shown to be probabilistically more significant [47, 50, 51].

(4) Spread: the distance between the first and last occurrences of the candidate concept, measured in terms of the number of characters and normalized by the length of the wiki page. This feature reflects the observation that a candidate concept which is more evenly spread within a tag's wiki page has a higher probability of being the tag's equivalent Wikipedia concept. In practice, this feature is expected to have a considerable impact only when the textual content of the tag's wiki page is of substantial length.

(5) Max Link Probability: the maximum value of the link probabilities of all the candidate concept's lexical forms which appear in the tag's wiki page. The link probability of a lexical form is the ratio of the number of times it occurs in Wikipedia articles as a hyperlink (directing to its corresponding article) to the number of times it occurs as plain text. This feature is based on the assumption that candidate concepts whose descriptor and/or non-descriptor lexical forms appearing in the tag's wiki page have a high probability of being used as a hyperlink in Wikipedia articles, would also have a high probability of being used as user tags.

(6) Average Disambiguation Confidence: in many cases a term in a tag's wiki page could correspond to multiple concepts in Wikipedia and hence needs to be disambiguated. For example, the term "Java" could refer to various concepts, such as "Java programming language", "Java Island", "Java coffee", etc. As described in [52], the Wikipedia-Miner uses a novel machine learning-based approach for word-sense disambiguation which yields an F-measure of 97%. We have set the Wikipedia-Miner's disambiguator to perform a strict disambiguation, i.e., each term in the wiki page can only correspond to a single concept which has the highest probabilistic confidence. The value of this feature for a candidate concept is calculated by averaging the disambiguation confidence values of its descriptor and non-descriptor lexical forms that appear in the tag's wiki page. This feature acts as a validity check mechanism for the detected concepts.

(7) Max Disambiguation Confidence: the maximum disambiguation confidence value among the lexical forms of a candidate concept which appear in the tag's wiki page. Both the average and max disambiguation confidence features are incorporated to reduce the equivalency likelihood score of those candidate concepts which have a low disambiguation confidence. A low disambiguation confidence value for a candidate concept sheds doubt on its existence and validity in the wiki page.

(8) Link-Based Relatedness to Other Concepts: the Wikipedia-Miner measures the semantic relatedness between concepts using a method called Wikipedia Link-based Measure (WLM). In this method the relatedness between two Wikipedia articles/concepts is measured according to the number of Wikipedia concepts which discuss/mention and have hyperlinks to both the two concepts being compared (see [22] for details). For example, "text mining" and "genetic algorithms" have 53% relatedness based on the fact that a third Wikipedia concept "artificial intelligence" has mentioned and has hyperlinks to both. The value of this feature for a candidate concept is obtained by measuring and averaging its relatedness to all the other candidates detected in the tag's wiki page. The tag's equivalent concept is expected to have a high semantic relatedness to the majority of other candidate concepts detected in the tag's wiki page, as together they form a cluster of related concepts each covering a specific aspect of the same subject/phenomenon discussed in the tag's wiki page.

(9) Link-Based Relatedness to Context: the only difference between this feature and the Link-Based Relatedness to Other Concepts is that the relatedness of the candidate concept is only measured against those of other candidate concepts in the tag's wiki page which are unambiguous, i.e., their descriptor and/or non-descriptor lexical forms occurring in the wiki page have only one valid sense. Both the Link-Based Relatedness to Context and Link-Based Relatedness to Other Concepts features are incorporated to increase the equivalency likelihood score of those candidate concepts which have a high semantic relevance to other concepts in the wiki page. However, the former only takes into account the unambiguous concepts in the wiki page and therefore has a high accuracy but low coverage, whereas the latter also includes the ambiguous concepts which have been disambiguated based on their surrounding unambiguous context (i.e., unambiguous concepts in the wiki page) and therefore has a lower accuracy but conclusive coverage.

To illustrate the application of the features defined above, consider the case of the RapidMiner concept which is the equivalent concept for the rapidminer tag used as an example in Section 3 (Figure 1). In this case, the RapidMiner concept has appeared in the tag's wiki page 5 times, which is the highest Frequency feature value among the other concepts detected in the wiki page, and therefore its normalized Frequency feature value is 1.0. This concept appears for the first time at the very beginning of the wiki page and therefore its First Occurrence feature value is 1.0. It reappears at the ending line of the wiki page for the last time and therefore it's Last Occurrence, and Spread feature values are 0.86 and 0.86, respectively. In this particular example, the Last Occurrence and Spread feature values are equal due to the appearance of the concept as the very first term in the wiki page. The Average Link Probability feature value of this concept is 1.0. This means that in all the cases where its descriptor or non-descriptor lexical forms have appeared in a Wikipedia article, they have been hyperlinked to the concept's main article in Wikipedia. The Average Disambiguation Confidence and Max Disambiguation Confidence feature values for this concept are both equal to 0.97, which means that in this case the disambiguator component of the Wikipedia-Miner has a very high confidence in the correctness of its disambiguation result. The Link-Based Relatedness to Context and Link-Based Relatedness to Other Concepts feature values for this concept are 0.53 and 0.78, respectively. At first glance, these feature values do not appear to be particularly high, however they are exceptionally high when compared to those of other candidate concepts detected in the wiki page. The RapidMiner concept has the highest Link-Based Relatedness to Other Concepts feature value among the other candidate concepts, and it's Link-Based Relatedness to Context feature value is the 4th highest after the concepts Data mining (0.60), Machine learning (0.57), and Predictive analytics (0.54). This may be interpreted as a possible early sign that the Link-Based Relatedness to Other Concepts is a more reliable feature than the Link-Based Relatedness to Context for our binomial classification task in this work. We have investigated this possibility in Section 5 using feature selection metrics.

4.2. Building a Training & Testing Dataset

Having defined a set of features for the Wikipedia concepts found in the tags' wiki pages, we then need to build a dataset of manually mapped StackOverflow Tag-to-Wikipedia Concept sample instances. This dataset is fed to a ML-based classification algorithm for learning a prediction model from. We also use the same dataset for evaluating the prediction accuracy performance of the classifier using a 10-fold cross-validation procedure (more on that in Section 5).

We adopted a semi-supervised labelling method to build our required dataset. This method consists of two main stages. In the first stage, we compile a set of sample tags which may be mapped to their corresponding Wikipedia concepts with a high level of accuracy using two simple rules:

(1) The tag's wiki page should contain one or more hyperlinks to Wikipedia articles.
(2) One of those hyperlinks should be to the Wikipedia article corresponding to the first Wikipedia concept detected in the tag's wiki page.

To illustrate, consider the case of the rapidminer tag, which we have been using as an example so far. As can be seen in Figure 1, the full description part of this tag's wiki page contains a hyperlink (appearing in bold blue font). This hyperlink is to the Wikipedia article for Rapidminer. Also, as already described in Section 4.1, the Rapidminer concept, which is the equivalent Wikipedia concept for the tag, appears at the very beginning of the tag's wiki page. Therefore, in case of this example both rules are satisfied. As the result, the Rapidminer concept is labelled as "equivalent" and all the remaining 15 concepts detected in the tag's wiki page are labelled as "non-equivalent". This mapping data is then added to the dataset.

We found a total of 1,256 tags which fulfilled the above requirements, and added their mapping data to the preliminary dataset. At this stage we were confident that, using above rules, the great majority of tags in the dataset were correctly mapped. Despite that, as the second stage of our semi-supervised labelling method, we manually inspected all the tags and their mapping data in the dataset to verify their correction and rectify any possible mis-mappings. As the result, we found a total of 16 mis-mappings (1.27%) caused by two types of errors: (a) the tag's equivalent Wikipedia concept is not among the concepts detected in the tag's wiki page; and (b) the tag's equivalent Wikipedia concept is detected in the tag's wiki page, but it is not the first detected concept. Table 2 shows the mis-mapped tags, their error types, wrongly mapped and true equivalent Wikipedia concepts. After rectifying these mapping errors, the final dataset contains a total of 1,256 tags, out of which 1,250 are correctly mapped to their equivalent Wikipedia concepts/articles and the remaining 6 tags are left unmapped as their equivalent Wikipedia concepts were not detected in their wiki pages (type (a) error cases). The final dataset contains a total of 38,184 Wikipedia concept instances, out of which 1,250 (3.27%) belong to the equivalent/true class and the remaining 36,934 (96.73%) concepts belong to the non-equivalent/false class. In other words, there is a 1:30 ratio between the equivalent and non-equivalent Wikipedia concepts detected in the wiki pages of the sample tags included in the dataset.

Table 2. Rule-based mapping errors.

StackOverflow Tag	Equivalent Wikipedia Concept	Wrongly Mapped Wikipedia Concept	Error Type
trust	wikipedia.org/wiki/computational_trust	wikipedia.org/wiki/Trust_(social_sciences)	(a)
collections	wikipedia.org/wiki/collection_(computing)	wikipedia.org/wiki/java_collections_framework	(a)
null	wikipedia.org/wiki/Null	wikipedia.org/wiki/Null_(SQL)	(a)
schema	wikipedia.org/wiki/Schema	wikipedia.org/wiki/database_schema	(a)
lambda	wikipedia.org/wiki/anonymous_function	wikipedia.org/wiki/lambda_calculus	(b)
windows-xp	wikipedia.org/wiki/Windows_XP	wikipedia.org/wiki/Microsoft_Windows	(b)
fork	wikipedia.org/wiki/Fork_(operating_system)	wikipedia.org/wiki/Fork_(software_development)	(a)
annotations	wikipedia.org/wiki/Annotations	wikipedia.org/wiki/Java_annotation	(b)
modularity	wikipedia.org/wiki/Modular_programming	wikipedia.org/wiki/Modularity	(b)
function	wikipedia.org/wiki/Function_(computer_science)	wikipedia.org/wiki/Function_(mathematics)	(b)
facade	wikipedia.org/wiki/Facade_pattern	wikipedia.org/wiki/Façade	(b)
gaussian	wikipedia.org/wiki/Gaussian_function	wikipedia.org/wiki/Normal_distribution	(b)
phone-number	wikipedia.org/wiki/Telephone_number	wikipedia.org/wiki/Telephone	(b)
prime-factoring	wikipedia.org/wiki/Prime_factorization	wikipedia.org/wiki/Prime_number	(b)
associativity	wikipedia.org/wiki/Operator_associativity	wikipedia.org/wiki/Associative_property	(b)
ext.net	wikipedia.org/wiki/Ext.NET	wikipedia.org/wiki/Ext_JS	(a)

5. Experimental Results & Evaluation

The dataset described in Section 4.2 is stored in Attribute-Relation File Format (ARFF)[25], which is the main file format used in Weka environment [53]. Weka is an open-source data-mining software tool, issued under the GNU General Public License, offering a comprehensive collection of data mining and machine learning algorithms. We have used Weka to experiment with and evaluate the accuracy performance of our proposed mapping method which uses an ML-based binomial classifier at its core. Table 3 shows the evaluation results of our experiments with various well-known ML-based classification algorithms and meta-algorithms, measured using standard information retrieval metrics and 10-fold cross-validation.

Table 3. Classification performance achieved using various classification algorithms and meta-algorithms in Weka.

Classifier (Weka implementation)	TP Rate	FP Rate	Precision	Recall	F_1	MCC	ROC Area	PRC Area
Bayes Network (BayesNet)	0.984	0.046	0.988	0.984	0.985	0.801	0.996	0.997
KNN (IB1 instance-based classifier)	0.986	0.219	0.986	0.986	0.986	0.779	0.890	0.980
SVM (LibSVM)	0.989	0.201	0.989	0.989	0.989	0.826	0.894	0.983
Decision Tree (J48)	0.996	0.065	0.996	0.996	0.996	0.931	0.974	0.994
Random Forest (RandomForest)	0.996	0.057	0.996	0.996	0.996	0.937	0.986	0.997
Bagging Random Forest	0.996	0.060	0.996	0.996	0.996	0.937	0.995	0.999
Random Committee Random Forest	0.996	0.064	0.996	0.996	0.996	0.937	0.996	0.999
Random Forest + Feature Selection all features except F9	0.996	0.060	0.996	0.996	0.996	0.935	0.988	0.997
Random Forest + Feature Selection all features except F2	0.989	0.195	0.988	0.989	0.988	0.815	0.978	0.994

TP, true positive; FP, false positive; MCC, Matthews correlation coefficient; ROC, receiver operating characteristic; PRC, precision-recall curve.

As can be seen in the results presented in Table 3, all the classification algorithms and meta-algorithms that we have experimented with have achieved an exceptionally high classification performance for our mapping task. The decision tree and random forest have achieved very similar results, outperforming the SVM, KNN, and Bayes network with a small margin in F_1 score (≤1.2%). Using meta-algorithms (bagging, random committee) did not result in a statistically significant improvement, which is to be expected considering the already very high F_1 score of 0.996 achieved by the random forest on its own. The high classification performances achieved here may partially be attributed to the high quantity and quality of the experimental dataset used. Using the semi-supervised labelling method proposed in Section 4.2, the compiled dataset is virtually noise free and contains a total of 38,184 labelled Wikipedia concept instances to train and test the classification algorithms with.

After establishing the performance of various classification algorithms for our mapping task, we then measured the effectiveness of each of the 9 features, defined for Wikipedia concepts in Section 4.1, using various feature selection metrics. For this purpose, we adopted four commonly-used feature selection metrics, namely Chi-squared, Info Gain, Correlation, and Symmetrical Uncertainty, which are all implemented in Weka. Figure 2 shows, in descending order, the average normalized ranks for each feature according to the above four feature selections metrics after 10-fold cross-validation.

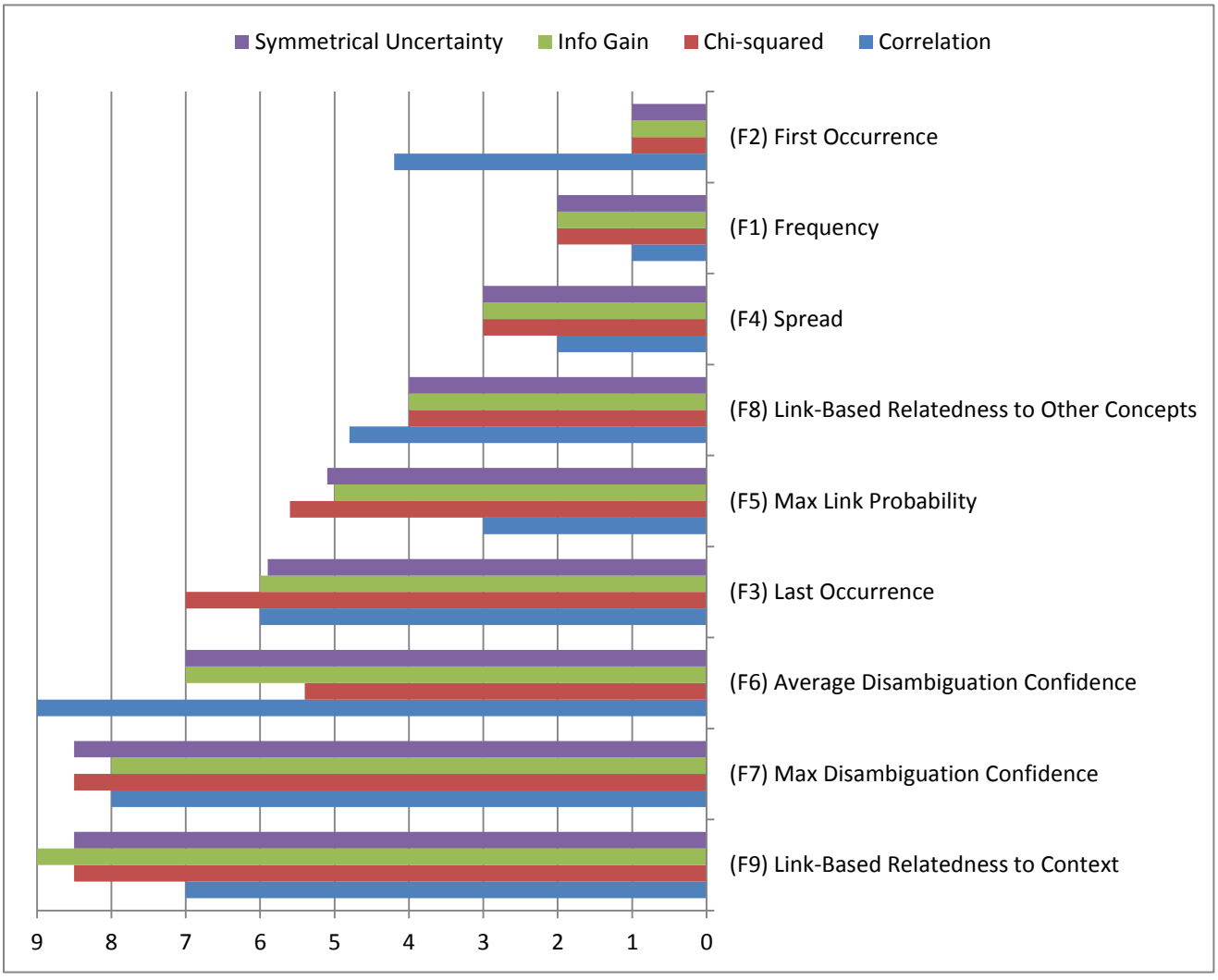

Figure 2. Average normalized ranks of the Wikipedia concept features according to four different feature selection metrics.

As shown in Figure 2, the 9th feature (F9), Link-Based Relatedness to Context, has achieved the lowest rank among other features, and therefore may be regarded as the weakest feature with the lowest positive impact on the accuracy

performance of the classification algorithms we have experimented with. We verified this assumption by re-training and testing the best preforming classification algorithm (i.e., Random Forest) on the dataset, but this time excluding the F9 feature. The second-last row of Table 3 shows the results of this test which confirms excluding F9 does not have a statistically significant impact on the overall classification performance. As discussed in the context of the example given at the end of Section 4.1, this may be attributed to the fact that F9 shows a bias towards more generic Wikipedia concepts detected in a tag's wiki page. Also, as speculated in Section 4.1, the higher ranking of F8 versus F9 confirms that the Link-Based Relatedness to Other Concepts is a more reliable feature than the Link-Based Relatedness to Context for our binomial classification-based mapping task in this work.

Looking at the other end of spectrum, F2 which is a positional feature capturing the first occurrence of a Wikipedia concept in a tag's wiki page, has come up as the strongest feature. This comes as no surprise, as we were already aware of, and counting on, the strength of this particular feature. As described in Section 4.2, our rule-based semi-supervised labelling method for compiling a training and testing dataset takes advantage of the high reliability of this feature to produce a preliminary dataset with an exceptionally low level of noise (only 1.27% mis-mappings). The above observation raised the question whether our semi-supervised labelling method has led to producing a skewed or biased dataset. If this is the case, the model learnt from the compiled dataset would be highly reliant on the existence and accuracy of a single feature, i.e., First Occurrence (F2), and therefore might not be able to achieve the same high levels of classification performance as those reported in Table 3, when used to map the rest of user tags in StackOverflow to their corresponding Wikipedia concepts. In order to investigate this possibility, we ran a final experiment in which we re-trained and tested the best preforming classification algorithm (i.e., Random Forest) on the dataset again, but this time excluding the F2 feature. The last row of Table 3 shows the results of this experiment. These results prove that even in the absence of the F2 feature, the classification model learnt based on the remaining features is capable of mapping the user tags to their equivalent Wikipedia concepts with a minimal loss of accuracy.

All the data used and generated in this work is available for download[26]. This includes: (a) an XML file containing all the user tags in StackOverflow and their corresponding usage frequency counts; (b) a log file containing the data produced during the process of detecting Wikipedia concepts in tags' wiki pages, and computing their feature values; (c) the manually verified StackOverflowTags-To-WikipediaConcepts-Mappings dataset in ARFF format, which may be readily used to duplicate all the reported experiments using Weka and to conduct further experimentation and analysis on.

6. Conclusion and Future Work

In this article, we described the design and development of a ML-based method for mapping user tags to their equivalent concepts in Wikipedia. In this context, Wikipedia serves as a comprehensive controlled vocabulary. The proposed mapping offers multi-facet benefits to the process of subject metadata generation and management in a wide range of online environments, where there is an abundance of user-assigned tags. Some of these benefits include:

- Normalization of inconsistencies in user-generated subject metadata due to the uncontrolled nature of user tags, and caused by spelling variations, synonyms, acronyms, and hyponyms. This in turn would eliminate some of the common problems associated with the use of folksonomies such as the tag explosion phenomenon.
- Elimination of personal tags (e.g., "to read", "unread") which have no subject indexing value for the community.
- Improving the interchangeability, integrability, and augmentability of existing subject metadata across different online platforms and environments.

In the proposed method, we first identify all the Wikipedia concepts appearing in the wiki page of the tag to be mapped as candidate target concepts. We then deploy a ML-based classification algorithm to classify the detected concepts into equivalent or non-equivalent categories. We showcased the application of the proposed mapping method and evaluated its performance using the currently most popular computer programming Q&A website, StackOverflow, as our test platform. We evaluated the performance of our method using the standard information retrieval metrics of precision, recall, and F_1. Depending on the ML-based classification algorithm used, F_1 scores as high as 99.6% were achieved.

We expect the encouraging performance results achieved in this experiment to be transferable to similar websites and content with little or no customization required. For example, the proposed method may be readily applied to any of the other 118 Q&A websites in the StackExchange network. Therefore, as future work, we plan to investigate the

application and performance of the proposed method in other domains. The performance of our mapping method in any domain is expected to highly depend on the coverage of that particular domain in Wikipedia. However, given the comprehensiveness of the English Wikipedia and its high rate of growth, we expect an acceptable level of performance to be achievable in many domains which are already well covered in Wikipedia. Another aspect of the proposed method which requires further investigation is the effect that the size of training dataset has on the performance of the ML-based binomial classifier used in our mapping task.

By mapping (effectively converting) user tags to Wikipedia concepts, the method proposed in this work paves the way for the design and development of new semantically enhanced Information Retrieval (IR) methods and systems for content classification, clustering, ranking, and recommendation in various online environments. To illustrate this point with the help of an example, consider the case of enhancing the IR capabilities and performance of the StackOverflow website by taking advantage of the new subject metadata produced by our mapping method. Having questions indexed with Wikipedia concepts allows us to measure the semantic similarity between questions using the Wikipedia Link-based Measure (WLM) approach (see [22] for details) with a considerably higher level of accuracy compared to that achieved by traditional TFIDF-based methods. This in turn can be used to improve the performance of clustering, ranking, and recommendation algorithms which rely on accurate calculation of such similarity measures at their core. Another example of such improvement would be the development of a new approach for automatic classicisation of StackOverflow questions according to the Wikipedia's extensive classification system. Considering the fact that over 95% of Wikipedia articles/concepts are already classified according to this system, the parent categories of any given concept may be considered as high-probability candidate parents for all the questions whose user-assigned tags are mapped to that concept. These candidate parent categories may then be used as reliable cues for automatic classification of StackOverflow questions according to the Wikipedia's classification system.

Finally, the proposed mapping could be utilized to initiate new links between Wikipedia articles/concepts and StackOverflow Q&As, where, for example, suitable Wikipedia articles could include lists of related Q&As from StackOverflow as further reading or related resources.

Notes

1. https://www.flickr.com/
2. https://vimeo.com/
3. https://www.librarything.com
4. http://www.citeulike.org/
5. http://www.ncbi.nlm.nih.gov/mesh
6. https://www.ncbi.nlm.nih.gov/pubmed/
7. http://scholar.google.com/
8. http://books.google.com/
9. http://stackoverflow.com/
10. https://delicious.com/
11. http://authorities.loc.gov/
12. http://www.oclc.org/dewey.en.html
13. http://wordnet.princeton.edu/
14. http://stackexchange.com/
15. http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
16. http://aims.fao.org/standards/agrovoc/faq#How often is AGROVOC updated?
17. http://id.loc.gov/authorities/subjects/sh96000740.html
18. http://en.wikipedia.org/wiki/Metadata
19. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53657
20. http://www.alexa.com/siteinfo/stackexchange.com
21. http://stackoverflow.com/tags/rapidminer/info
22. https://archive.org/details/stackexchange
23. https://archive.org/download/stackexchange/readme.txt
24. http://en.wikipedia.org/wiki/RapidMiner
25. http://www.cs.waikato.ac.nz/ml/weka/arff.html
26. http://www.skynet.ie/~arash/zip/StackOverflow_Wikipedia_May2014.zip

Funding

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

References

[1] Mathes A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata, http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html (2004, accessed March 2015).
[2] Trant J. Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information*. 2009; 10.
[3] Guy M and Tonkin E. Folksonomies: Tidying up Tags? *D-Lib Magazine*. 2006; 12.
[4] Caimei Lu, Park J-r and Xiaohua Hu. User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*. 2010; 36: 763-79.
[5] Lee DH and Schleyer T. Social tagging is no substitute for controlled indexing: A comparison of Medical Subject Headings and CiteULike tags assigned to 231,388 papers. *Journal of the American Society for Information Science and Technology*. 2012; 63: 1747-57.
[6] Wu D, He D, Qiu J, Lin R and Liu Y. Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese. *Journal of Information Science*. 2013; 39: 169-87.
[7] Cantador I, Konstas I and Jose JM. Categorising social tags to improve folksonomy-based recommendations. *Web Semantics: Science, Services and Agents on the World Wide Web*. 2011; 9: 1-15.
[8] Cantador I, Szomszor M, Alani H, Fernández M and Castells P. Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations. *1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008)*. Tenerife, Spain: Springer, 2008.
[9] Noruzi A. Folksonomies: Why do we need controlled vocabulary? *Webology*. 2007; 4.
[10] Wikipedia. Wikipedia:Size in volumes, http://en.wikipedia.org/wiki/Wikipedia:Size_in_volumes (2014, accessed Oct 2014).
[11] Rainie L and Tancer B. Wikipedia users, http://www.pewinternet.org/Reports/2007/Wikipedia-users.aspx (2007, accessed July 2014).
[12] Yi K and Chan LM. Linking folksonomy to Library of Congress subject headings: an exploratory study. *Journal of Documentation*. 2009; 65: 872-900.
[13] Golub K, Lykke M and Tudhope D. Enhancing social tagging with automated keywords from the Dewey Decimal Classification. *Journal of Documentation*. 2014; 70: 801-28.
[14] Angeletou S, Sabou M and Motta E. Semantically enriching folksonomies with FLOR. *1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008) at The 5th Annual European Semantic Web Conference (ESWC 2008)*. Tenerife, Spain2008.
[15] García-Silva A, Corcho O, Alani H and Gómez-Pérez A. Review of the state of the art: discovering and associating semantics to tags in folksonomies. *The Knowledge Engineering Review*. 2012; 27: 57-85.
[16] Medelyan O, Milne D, Legg C and Witten IH. Mining meaning from Wikipedia. *Int J Hum-Comput Stud*. 2009; 67: 716-54.
[17] Bunescu RC and Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation. *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*. Trento, Italy2006, p. 9-16.
[18] Wang P and Domeniconi C. Building semantic kernels for text classification using wikipedia. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Las Vegas, Nevada, USA: ACM, 2008, p. 713-21.
[19] Hu X, Zhang X, Lu C, Park EK and Zhou X. Exploiting Wikipedia as external knowledge for document clustering. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris, France: ACM, 2009, p. 389-96.
[20] Ciglan M and Nørvåg K. WikiPop: personalized event detection system based on Wikipedia page view statistics. *Proceedings of the 19th ACM international conference on Information and knowledge management*. Toronto, ON, Canada: ACM, 2010, p. 1931-2.
[21] Medelyan O, Witten IH and Milne D. Topic Indexing with Wikipedia. *first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*. Chicago, US: AAAI Press, 2008.
[22] Milne D and Witten IH. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08)*. Chicago, I.L2008.
[23] Zesch T, Gurevych I, M M, et al. Comparing Wikipedia and German wordnet by evaluating semantic relatedness on multiple datasets. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Rochester, New York: Association for Computational Linguistics, 2007, p. 205-8.
[24] Milne D, Medelyan O and Witten IH. Mining Domain-Specific Thesauri from Wikipedia: A Case Study. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 2006, p. 442-8.
[25] Vivaldi J and Rodríguez H. Finding Domain Terms using Wikipedia. *Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010.
[26] Osborne M, Petrovic S, McCreadie R, Macdonald C and Ounis I. Bieber no more: First Story Detection using Twitter and Wikipedia. *SIGIR Workshop in Time-aware Information Access (TAIA'12)*. Portland, Oregon, USA: ACM, 2012.
[27] Wikipedia. List of Wikipedias, http://en.wikipedia.org/wiki/List_of_Wikipedias (2014, accessed September 2014).
[28] Melo Gd and Weikum G. MENTA: inducing multilingual taxonomies from wikipedia. *Proceedings of the 19th ACM international conference on Information and knowledge management*. Toronto, ON, Canada: ACM, 2010, p. 1099-108.

[29] Salah AA, Gao C, Suchecki K and Scharnhorst A. Need to Categorize: A Comparative Look at the Categories of Universal Decimal Classification System and Wikipedia. *Leonardo*. 2012; 45: 84-5.
[30] Callahan ES and Herring SC. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*. 2011; 62: 1899-915.
[31] Giles J. Internet encyclopaedias go head to head. *Nature*. 2005; 438: 900-1.
[32] Britannica. Fatally Flawed - Refuting the recent study on encyclopedic accuracy by the journal Nature, http://corporate.britannica.com/britannica_nature_response.pdf (2006, accessed July 2014).
[33] Xu G, Wu Z, Li G and Chen E. Improving contextual advertising matching by using Wikipedia thesaurus knowledge. *Knowl Inf Syst*. 2014: 1-33.
[34] Macías-Galindo D, Wong W, Cavedon L and Thangarajah J. Using a Lexical Dictionary and a Folksonomy to Automatically Construct Domain Ontologies. In: Wang D and Reynolds M, (eds.). *AI 2011: Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2011, p. 638-47.
[35] Ponzetto SP and Strube M. Deriving a large scale taxonomy from Wikipedia. *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*. Vancouver, British Columbia, Canada: AAAI Press, 2007, p. 1440-5.
[36] Fogarolli A. 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, 2011, p. 1-26.
[37] Milne DN, Witten IH and Nichols DM. A knowledge-based search engine powered by wikipedia. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. Lisbon, Portugal: ACM, 2007, p. 445-54.
[38] Singer L, Filho FF, Cleary B, Treude C, Storey M-A and Schneider K. Mutual assessment in the social programmer ecosystem: an empirical investigation of developer profile aggregators. *Proceedings of the 2013 conference on Computer supported cooperative work*. San Antonio, Texas, USA: ACM, 2013, p. 103-16.
[39] Stack Exchange statistics, http://stackexchange.com/sites?view=list#traffic (2014, accessed September 2014).
[40] Parnin C and Treude C. Measuring API documentation on the web. *Proceedings of the 2nd International Workshop on Web 20 for Software Engineering*. Waikiki, Honolulu, HI, USA: ACM, 2011, p. 25-30.
[41] Barua A, Thomas S and Hassan A. What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empir Software Eng*. 2012: 1-36.
[42] Mamykina L, Manoim B, Mittal M, Hripcsak G and Hartmann B. Design lessons from the fastest q&a site in the west. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Vancouver, BC, Canada: ACM, 2011, p. 2857-66.
[43] Nasehi SM, Sillito J, Maurer F and Burns C. What makes a good code example?: A study of programming Q&A in StackOverflow. *Software Maintenance (ICSM), 2012 28th IEEE International Conference on*. 2012, p. 25-34.
[44] Treude C, Figueira Filho F, Cleary B and Storey M-A. Programming in a socially networked world: the evolution of the social programmer. *In FutureCSD '12: Proceedings of the CSCW Workshop on the Future of Collaborative Software Development*. 2012.
[45] Meier W. eXist-DB, http://exist.sourceforge.net/ (2014, accessed February 2014).
[46] Milne D. An open-source toolkit for mining Wikipedia. *New Zealand Computer Science Research Student Conference*. 2009.
[47] Medelyan O. Human-competitive automatic topic indexing. *Department of Computer Science*. University of Waikato, New Zealand, 2009.
[48] Hulth A. Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction. *Department of Computer and Systems Sciences*. Stockholm University, 2004.
[49] Turney PD. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*. 2000; 2: 303-36.
[50] Witten IH, Paynter GW, Frank E, Gutwin C and Nevill-Manning CG. KEA: practical automatic keyphrase extraction. *fourth ACM conference on Digital libraries*. Berkeley, California, United States: ACM, 1999.
[51] Mahdi AE and Joorabchi A. A Citation-based approach to automatic topical indexing of scientific literature. *Journal of Information Science*. 2010; 36: 798-811.
[52] Milne D and Witten IH. Learning to link with wikipedia. *Proceedings of the 17th ACM conference on Information and knowledge management*. Napa Valley, California, USA: ACM, 2008, p. 509-18.
[53] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten IH. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*. 2009; 11.