Providing suitable suggestions from online Q&A communities during Brainstorming

Saviour Agius
Department of Intelligent Computer Systems
Faculty of Information and Communication Technology
University of Malta
sagi0021@um.edu.mt

ABSTRACT
The “undiscovered knowledge problem” refers to the set of possible discoveries which one could realise had her knowledge span over several areas, thus being in a position to infer innovative findings. While this lacks, individuals prefer to excel in one or in a small number of specialised areas rather than trying to acquire knowledge across vast fields of study [16]. By exploiting the expertise of others, individuals may gain insight into additional branches in their field of interest; yet on the other hand they may also realise innovative understanding in fresh realms.

In this paper we combine a brainstorming component and take advantage of an online Q&A repository – to acquire relevant data in relation to a user’s idea or uncertainty – the proposed prototype is intended to produce a set of suggestions which effectively are the result of a term extraction process. Answers posted by highly reputed individuals, will act as the foundation for the formulation of key phrases which will eventually be presented back to the user in the brainstorming activity as suggestions.

1. INTRODUCTION AND BACKGROUND
Human interaction and sharing of ideas is a means through which individuals are able to procure knowledge, be it in a domain of interest or fresh ones – as is the aim of a brainstorming activity. The core aspect behind such an approach is to effectively come up with promising new ideas in order to solve a problem at hand. As promising as this may seem, brainstorming activities suffer as a consequence of member’s inability to recall specifics and if no proficient notions are contributed, such knowledge-achieving technique risks losing its worth.

Help-seeking communities such as Q&A forums present a subset of the forum sphere; providing a place in which casual users are able to elaborate their issue to expert users in view of obtaining insightful human feedback and directives based on expertise and past experiences [10]. Typically, forums provide reputation mechanisms through which a user’s virtual profile is credited for replying or posting questions which peer users deem to be relevant or appropriate. Identifying an individual or a cluster of individuals capable of providing adequate responses to a given question for some topic domain is tasked as the role of Expertise Finders [22, 23]. Material posted by knowledgeable persons is better appreciated and easier to trust [19]. Expert finding systems can – according to [22] – be considered as “a class of recommender systems, suffering from an inability to infer expertise levels”.

Recommender systems pose a special form of information filtering systems [13] in that the sole purpose behind the use of a recommender system is to provide meaningful suggestions to users for items of interest [12]. The two forms of recommender systems which are most prevalent include Collaborative Filtering (CF) and Content–based (CB) filtering [11]. A hybrid approach is a third viable option through which the benefits of both are combined into one to prevail over limitations which the two pure recommender flavours possess. When the Collaborative Filtering approach is employed, users are recommended items based upon past ratings of all users; conversely, in a content–based approach a system recommends items, whose content is similar to others a user has already liked in the past [11].

Keyphrases are intended to convey a brief notion from a bigger context; used to manage large amounts of contextual data [17] and are the result of a Term Extraction process. Automatically extracting adequate expressions and key-phrases within the forum domain, has proven to be a difficult task; work discussed in [3, 9] falls short in offering a solid approach, and shift the task of tackling the content aspect to human judges. An Automatic Term Recognition routine is analogous to that required in Information Extraction, where the process attempts to haul out salient data present in a number of referenced documents [5], achieved over two stages. A linguistic analysis which identifies candidate terms within a document is followed by a statistical task to determine the relevant importance of the extracted terms within the same document or corpus [1].

We propose to surpass the limitation outlined in brainstorming activities through the inclusion of a recommendation component allowing users to benefit from the knowledge explicitly demonstrated by proficient users in their attempt to respond to questions posted by fellow, less proficient members on an online Q&A community. The recommendations provided will be the result of single and multi-word term extraction algorithms applied on the answers.

The rest of the paper is organised as follows. In the next section, we discuss the design issues with regard to our work; the prototype implementation is discussed next. In section 4, we present our evaluation and remarks followed future work prospects. The paper comes to an end with a conclu-
sion as regards the efforts and achievements in relation to our efforts.

2. DESIGN

Many of the efforts in the domain of forums have been devoted either to the identification of expert users [10, 22, 4, 19, 23] or otherwise address the subject of answer quality [3, 9]. To the best of our knowledge, research which attempts to spark new ideas through content analysis whilst taking the replier’s expertise into consideration is still lacking and thus, our effort can be considered as innovative.

The aim of this work is to have a system which provides users with adequate suggestions during a brainstorming activity to generate new notions in relation to a current user problem. The ideal suggestions should be in the form of key-phrases as a result of an extraction process from the answers of the most proficient users within a Q&A community, on the premise that answers posted by highly reputed users constitute the best subject matter to provide adequate recommendations from. A second component involves the construction of a dataset from Q&A sets to model the communication flow which effectively demonstrates the expertise of individuals, to distinguish between common and proficient users.

The idea behind the Forum Module is that to utilise mined knowledge stored in active online repositories; to realise the objective of finding a number of experts required for a topic within a Q&A community, it will comprise of various algorithms which assess the text in one of two ways. This analysis will either be a surface-based process based on frequency of terms, or alternatively consider a Natural Language approach, thus employing a more comprehensive methodology.

In the field of Automatic Term Recognition (ATR), techniques often relied on either a statistical or a linguistic approach; however, recent ones considered a hybrid approach [1, 7, 2]. This merely advocates that a unified approach, incorporating both a statistical and a linguistic analysis is the way forward. Hence, due to such motivation, we will confront the problem at hand from a somewhat surface-based aspect through frequencies of terms, a unified approach involving the hybrid of a linguistic and statistical approach, along with a deeper breakdown and assessment of the text through Natural Language Processing.

The JATR library encompasses a subset of algorithms within the ATR domain identified and implemented by [24] in which the performance of each algorithm was then compared on a corpus from Wikipedia⁶ and another one from GENIA². The consideration of the Xtrak4Me³ and the JATR⁴ libraries stems from the notion to confront the problem from the aspects previously outlined and to investigate their enactment within our system. The TF-IDF and Weirdness algorithms provide a somewhat surface-based aspect to the analysis of answer text through term frequencies; the C-Value algorithm involves a more comprehensive analysis through the hybrid approach. JATR was also selected on the grounds that it implements a voting mechanism by combining the results from the different methods (single and multi-word terms) into an integrated output; expected to improve the results of the joint methods when considered on a separate basis. The Xtrak4Me platform on the other hand makes use of several GATE⁵ components to accomplish the automatic extraction of keyphrases whilst allowing developers to do so in an efficient manner.

KEA⁶ was also considered however such library can only be utilised if supplied with text documents. Moreover, the complexity of manually building a model for every document without knowing what the contents are, was deemed as an additional burden. Taking into account how Xtrak4Me provides developers with a straightforward utilisation procedure, whilst still employing GATE technologies as an underlying mechanism, this library seems to fit the bill quite well.

2.3 The Expertise Module

To realise the objective of finding a number of experts related to a particular area we intend to construct a dataset which can then be utilised as a repository to analyse the flow of communication between users. The more replies an individual poses, the more likely are the chances that that same individual is knowledgeable about the topic. The motivation behind this aspect, stems from research presented

2.http://www.nactem.ac.uk/genia/
5.http://gate.ac.uk/
in [22, 23, 10] compiled in online Q&A forums, in addition to finding experts in an academic researcher network as in [21]. The authors come up with an asker-helper network by mapping a subset of users who ask a question to those that respond to the question; considering the repliers as having a better understanding – thus better expertise – about the topic.

The dataset is intended to comprise several Q&A pairs – including various attributes to both questions and answers – extracted from the Stack Exchange network which could serve as the basis for such construction. We propose that the dataset be developed in GraphML. This is an XML-based file intended to describe the structural properties of graphs in a universal format. GraphML supports several graph structures including directed, undirected and mixed graphs, in addition to hyper-graphs and application-specific attribute data. It is also supported by a considerable number of graph analysis tools, such as Gephi and JUNG.

3. IMPLEMENTATION

The prototype is implemented in Java as this facilitated the use of the term extraction libraries discussed in this section. The subsequent subsections will illustrate each component in better detail.

3.1 The Forum Module

This can be considered as the primary building block of our system through which we gain access to an online knowledge repository from which information is extracted and operated on, in subsequent processes. We implemented the IForum and IForumAnswers interface as per the needs for the Stack Exchange API. The Stack Exchange network was selected due to its wide-range of forums available and due to its extensive audience and our work focuses on two particular forums – the sports and biology forums. The sports forum has been selected mainly due to the utter presence of text in answers as opposed to stack overflow, where answers could be simply constituted of code, and thus eliminate the possibility of providing any form of suggestions. The inclusion of the biology domain stems from the consideration that ATR approaches have been widely evaluated in biological domains.

The primary functionality we required for our application was the search facility to which we provide a forum name and a string containing the text which is then matched to a number of relevant questions. A filter function, presented us with the possibility of altering the attributes of the default response object returned for a given request. Thus we were able to obtain all the relevant information attributes to both the questions matched and their answers in a single request; this was pertinent so that we keep to a bare minimum the number of requests issued and consequently do not undermine system efficiency.

3.2 Keyword Extraction

This module encapsulates the heart and soul of our prototype, with both the Xtrak4Me and the JATR platforms producing keyphrases and an associated coefficient. Each set of results produced by each platform is stored in a HashMap<String, Double> so as to store the extracted terms together with their respective weight.

3.2.1 Xtrak4Me

The work compiled by Schutz [17] is freely accessible as a Java application. The library comes bundled with a number of GATE plug-ins which must be referenced before the Keyphrase extractor is executed. Once initialised, the component is able to extract keywords and keyphrases from URL and Strings. As question answers are what effectively form our corpus, these are what we supply for extraction.

3.2.2 JATR

This library is the work of [24] which comprises of a number of Automatic Term Recognition (ATR) algorithms capable of handling single and multi-word terms. We make use of the available JAVA library, bundled with a number of corpora on which every ATR algorithm can be applied to. Such corpora are rather insignificant for us and consequently we developed a work-around so as to keep the notion of a corpus, but alter the subject matter. An intermediary class, AnswerDocument – which implements the Document Interface defined by the JATR library – has been developed to handle such aspect; given an instance of an Answer object, a Document instance is created and consumed by the JATRUtil class when building a corpus.

The JATR library is bundled with a number of corpora on which every ATR algorithm must be applied to which to us is rather insignificant. For this reason, we developed a work-around so as to keep the notion of a corpus, but alter the subject matter. An intermediary class, AnswerDocument – which implements the Document Interface defined by the JATR library – has been developed to handle such aspect; given an instance of an Answer object, a Document instance is created and consumed by the JATRUtil class when building a corpus.

The JATRUtil class handles utility methods such as the building a corpus in relation to a question, constructing a Map<String, Double> from a Term array, and obtaining a reference to the BNC file bundled with the JATR library. BNC is short for the British National Corpus and provides a vast collection of written and spoken words devised as to characterise a wide sample of English words. The BNC is solely used by GlossEx, TermEx and Weirdness.

The JATRExecutor mainly consists of an execute method which must be supplied with a List<Answer>. This method must be implemented by all classes that extend the AbstractJATRAlgorithm abstract class, as well as by the VotingJATRAlgorithm. The reason we adopted such structure lies in the fact that the Voting algorithm necessitates all the other ATR algorithm to provide a score for each term; whilst its underlying functionality differs in contrast to the
The intent behind the AbstractJATRAlgorithm was to act as a compromise with the JATR library, its ATR algorithms, and their employment within our prototype. The development of this abstract class diminishes any future efforts involved to incorporate new JATR algorithms for any custom corpus.

3.3 Visualisation Media
After generating suggestions, the appropriate means to convey the end result must follow. We believe that different users are accustomed to different representation mediums, and what one may consider as good, it may be of no benefit to others. Hence for such reasons, users using the prototype are presented with three different result formats, conveniently assembled within a CardLayout. A Tag Cloud, a ranked list, and the actual answers of which forum members provided to the asker of the question are all combined in a SuggestionsPanel.

Tag Cloud. One of our initial objections was to present users with meaningful suggestions from answers having been passed through a term extraction process; the inclusion of a tag cloud to represent the salient terms seemed utterly appropriate. Our aspiration was realised through the Open-Cloud\textsuperscript{14},\footnote{http://opencloud.mcavallo.org/} Java library, which allowed us to supply word and value pairs, which essentially constitute the terms to be generated; the value is then used to determine the textual magnitude of each term. In assigning terms to a tag cloud, Open-Cloud consents developers to associate a web-link with each term. This functionality was exploited to provide users with the possibility of obtaining more information as regards the keywords by directly linking a term to a Wikipedia\textsuperscript{15} page – if existent.

A Ranked List. In addition to representing our results in a tag cloud format, we felt that representing the results in a ranked list also seemed relevant. Thus to illustrate the results, we employed Java’s DefaultTableModel in order to populate a JTable through the Map containing the terms and their associated weight. An intermediary step to normalise the values within the results map was enforced; this regulated the values into a between 0 and a 100 for all term extractor algorithms, as the values produced where not within a pre-defined range. Moreover, a threshold value of 35 was set so as to filter out non-relevant terms; this threshold value was determined after a number of tests were carried out and the relevant suggestions were assigned a superior value on a regular basis.

The Actual Answers. Another alternate way of depicting suggestions, was by providing the actual answers supplied by the users on the forum to the questions that matched the query issued by the brainstormers in our system. Attributes associated with an answer such as the username of the replier as well as the replier’s reputation and whether the answer was the accepted one or not are presented.

3.4 Modelling of Expertise
The idea here was to produce a dataset of Q&As from which one could identify the expertise of individuals through the flow of communication. Such information was to be stored in GraphML where a node represents users together with other attributes associated with a user, such as: username, id, reputation and question or answer id – depending upon whether the user was a replier or an asker. Alternatively the edges constituted whether the answer – posted by a user from whom the arrow points away – is an accepted answer or not, together with the score associated to that question. It goes without saying that a directed graph structure was employed for this scenario.

Through the utilisation of a graphical analysis tool, such as JUNG or Gephi, one can model and analyse graphs through visualisations. Furthermore such tools provide network analysis algorithms such as HITS and PageRank which could be exploited to obtain a relative user score within the dataset. Such aspect was however partially implemented and thus we cannot offer tangible conclusions as regards this component.

4. EVALUATION
We consider this stage as the most important of our research, hence we made sure to tackle all ambiguity well. First we look at some literature and evaluative techniques employed in the field of Term extraction and Recommender systems; then we put forward our evaluative approach prior to presenting the evaluation results and remarks provided.

4.1 Term Extraction Evaluation Research
Zhang et al.\textsuperscript{[24]} draw attention to the fact that evaluation research in the domain of term recognition is uneven, consequently implying that no standard approach exists. Vari-ety of techniques involve corpus size, domain selection, and evaluation mechanisms such as human judges or a dictionary based approach. The writers put the algorithms within the JATR library under even experimental conditions to compare and assess their performance on a collection of texts from Wikipedia, and a selected corpus from GENIA. The authors substantiate such selection by stating that several ATR approaches have been evaluated in either a medical or a biological domain, and solicit whether terms in such fields are more domain-specific.

A different number of evaluation approaches adopted in the field of ATR and even though they tackle the problem from a different perspective, only analyse a subset of the results. Human judges have been employed to manually assess a selected portion of the output\textsuperscript{[7, 15, 18]}, whilst in\textsuperscript{[6, 20]} an unsupervised matching of sections of the output against a dictionary source where performed. Zhang et al. employ 3 human judges to manually inspect the top 300 candidate terms that each algorithm produced, marking those they believe to be terms one would expect to encounter when reading texts about animals.

4.2 Recommender Systems Evaluation Research
The research by\textsuperscript{[8]} focuses on the development of a common framework through which recommender systems can
be compared, and put forward a number of evaluation metrics.

**Accuracy metrics.** Described as the degree of *nearness* to the true value achieved by the system, commonly used in the field of Artificial Intelligence, and aptly modified to accommodate such domain (see equation (1)). A second widespread evaluation metric includes the MAE – Mean Absolute Error – representing the average absolute deviation of every predicted rating against a user’s real attributed rating.

\[
\text{accuracy} = \frac{\text{number of successful recommendations}}{\text{number of recommendations}}
\]  

(1)

**Information retrieval measures.** Information Retrieval measures are also adapted to the realm of Recommender Systems in a similar manner as to the accuracy measure. The writers duly point out that Information Retrieval deals with the retrieval of relevant documents from some collection; this characterisation is mapped onto the Recommender System domain as the problem of recommending useful items from a pool. As precision and recall measures are vastly employed within the domain of Information Retrieval, and as such domain conveys influence over Recommender Systems, both measures are also used as evaluation means for Recommender Systems.

**F-measures.** The last metric discussed include the F-measures, portrayed as derivatives from the precision and recall measures, and that “try to grasp” the behaviour of the precision and recall metrics in a single value. The most popular of the F-measures is the F1 measure, known as the harmonic mean of precision and recall:

\[
F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]  

(2)

4.3 Proposed Evaluation Framework

Based upon the insight offered by [24] we presented our prototype to a number of participants for manual assessment. However, instead of just simply presenting the solution to a number of individuals, it was ensured that the participants possess a good understanding of the domain. This applies particularly for the evaluation of the biology forum – a specialisation area – where each of the selected participants was asked whether she was in possession of an “A” Level in Biology. Similarly, however less rigorous, participants for evaluating suggestions from the sports forum were expected to have a good understanding of the domain.

By presenting our solution to knowledgeable individuals, we aim to assess the suggestions presented, from the point of view of skilled participants; as opposed to individuals who will use our solution as a means of acquiring ideas – even though this is the scope of the prototype. However, we consider adept participants to be in a better position to distinguish between relevant and unrelated recommendations, and given their knowledge within the domain, feedback is expected to be more accurate.

4.4 Instruction to candidates

Participants were instructed to visit the respective forum site pertaining to the Stack Exchange network, familiarise themselves with the forum whilst executing a number of questions. Subsequently they were asked to select a question and present it to the prototype together prior to selecting an answer type and a term extraction algorithm. Every execution of the prototype resulted in three recommendation formats: a Tag Cloud, a ranked list and the replication of the actual answers posted by forum members. The selected candidates where instructed to run each algorithm on every answer type and rate the suggestions on a scale of one to five, before moving onto the next – a process which on average took around 30 minutes.

4.4.1 Discussion of the results

The jats:properties file within the JATR library stipulates a maximum length of candidate terms with a default value of 5. After a series of test runs during the evaluation process, it was noted that the C-Value algorithm was capable of yielding phrases composed of 8 words in length. For such reason this value was augmented to a round 10.

The most successful of all has been the C-Value due to its specificity to multi-word terms, whilst outperformed its contenders by quite some margin as was also highlighted by the participants themselves through their remarks. Employing a linguistic and statistical combination, it was capable of ousting XtraK4Me – a platform which makes use of GATE to analyse the corpora supplied and being employed in a number of studies for extraction purposes – by more than 50%.

The Voting algorithm, placing a third just after XtraK4Me did on occasions provide average results, however most of the time, its performance was considered below par. Although this is in-line with the findings of [24] (as it placed second to the C-Value algorithm), the Voting algorithm was expected to provide better outcomes especially when taking into consideration that it employs the other JATR algorithms as its underlying mechanism in an attempt to improve results [24].

The TFIDF was limited in its usefulness due to the answer types. Having a single answer meant having a single corpus and this essentially nullified the second part of its equation, effectively producing no results. On the positive side however, this algorithm was capable of providing multi-word terms on numerous occasions (on both forums) when suggestions were carried out from all of the answers pertaining to a single question.

GlossEx, TermEx and Weirdness account for the remaining three algorithms. TermEx produced satisfactory results on the Wikipedia corpus in [24], however it did not perform too well on the GENIA corpus. As regards both our domains, it fell short of providing adequate results on the sports forum, though it was expected to do well for the same reason it did on the Wikipedia corpus. Similarly can be said for the GlossEx and Weirdness algorithms, whose main employment within the system was to compare their performances against the other algorithms as well as to serve as a foundation for the Voting algorithm.
5. FUTURE WORK
A number of possibilities for future work were influenced by the current limitations as well as ideas that emerged from additional research whilst the implementation was in its later stages and no adequate implementation time was available.

The Brainstorming Application. We will look to provide users with a mechanism to rate suggestions, as well as a way to maintain the details of the positive suggestions – such as the forum member who posted the answer. These could be used for future inquiries, and answers by the same replier would be given preference.

Also, the creation and maintenance of a user profile provides an intriguing aspect. Had users accept to integrate their online account(s) on the Q&A forums being utilised by the application, the system could make use of a user’s past history (e.g. questions asked and answers provided) and keep a record of the her activity. If the application had to be used by a substantial amount of users, suggestions provided to one individual could be also presented to another user – provided that the asking and questioning history of both users is within a reasonable similarity score. This would result in a form of a collaborative filtering recommender system. Such initiative stems from the work by [14] where the Facebook account of the users was incorporated to extract interests for music, movies and hobbies; however we adapt its scope.

Term Extraction Mechanisms. Research in this area was a continuous effort, however, the evaluation results suggest a reassessment of the technologies employed, at least with regards to our area. The C/NC-Value method of term recognition may have been a better candidate than most of the JATR components. Such approach can be considered as an add-on to the C-Value method, in that it provides a more accurate coefficient by taking into consideration contextual cues, as stated by [2].

Another aspect which also raised a red flag was the effectiveness of the XtraKt4Me in the keyphrase extraction process. Having been employed and evaluated in a medium-size corpus as well as a practical setting, with state-of-the-art evaluation results [17], our findings can be considered as rather unsatisfactory. For this reason, in the future we consider making direct use of GATE and its underlying NLP mechanisms.

Expert Finding. The point of departure should include the completion of the Expertise Module as outlined in section 2.3. Though some implementation in this regard has been done this is still a long way from the initial idea. First off, the GraphML dataset should to be completed to act as the foundation for such task. Through the dataset, the ID of an asker can be linked to that of a replier and in order to produce an asker-helper network, similar to work by [22, 23, 10], in which an analogous approach was taken on a subset of the data from the Yahoo! Answers and Java Forums. By employing link-analysis algorithms such as PageRank, HITS and Z-score we can provide a relative rank of the users within the dataset. If a user is capable of answering questions of other individuals, it stands to reason that the replier has greater expertise than the asker; hence, a higher network score suggests a superior expertise level.

Evaluation Methodologies. As our current evaluation methodology adheres closely to the approach applied by [24] in that we as discussed in section 4.1, further evaluation methods could compliment the techniques described in section 4.2 to assess the recommendation aspect of our solution from a different perspective.

Instead of simply asking participants to rank algorithms based on what they believe is the production of rightful terms, we would instruct the candidates to first inspect the answers and put on paper a number of phrases or keywords they feel that would likely be the most indicative to users. As a side note, one may take into account that a future version of the current prototype, may employ other term extraction mechanisms. Nevertheless, once the participants complete their initial assignment, the term extraction algorithms should be executed in a similar manner as to the current ones. These are then followed by an evaluation process involving the metrics outlined in section 4.2.

6. CONCLUSION
Our goal was to explore a proof of concept which allows a user to augment his understanding with regards to a current problem at hand through the assistance of feasible recommendations in a brainstorming activity. We started off by delving into research that relates to our objectives namely the fields of Brainstorming, Expert Finding, Answer Quality and most importantly the area of Term Extraction – a subdivision of the Information Extraction discipline.

As evaluation remarks signified, not all algorithms provided the satisfactory results one would have hoped for. Nevertheless, the effects of such drawback was diminished as a result of the replication of the actual results pertaining to the question. Though several of the algorithms provided single-word terms – which on their own are not that suggestive – when participants refered to the actual answers, they were in a better position to put into context such suggestions. Quite a number of the chosen candidates expressed a positive acknowledgement as regards the selection of the visualisation media. Other than the replication of the answers, these involved a ranked list of the terms, and a tag cloud in which the generated terms were linked to Wikipedia entries for further clarification, keeping in mind that the goal of such system is to procure knowledge.

Overall, we are satisfied with what has been achieved through this work, given the time and resources at our disposal. Good results for the term extraction process were in the minority, but in spite of this, we were still able to realise an average to good score in relation to our final solution. We are confident that this latter aspect can be further improved by future research and development as to be capable to influence the domain of brainstorming and knowledge gain in the real world.
7. REFERENCES


[14] Jai Ramanarayanankrishnaniye Xinyu Xing Qin Lv Richard Han Shivakant Mishra Karim Saeda Mike Gartrell, Aaron Beach. Integrating wikipedia, facebook, and other personal online context into collaborative e-brainstorming. 2010.


