Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 6; April-June, 2015 pp. 542-545

http://www.krishisanskriti.org/acsit.html

Quality Prevision for Knowledge Base through CQUAL

Aatif Jamshed[1], Samiksha Manglani[2], Sundaram Raghuvanshi[3] and Trishla Jaiswal[4]

[1]Galgotias College of Engg. and Tech., Greater Noida
[2,3,4]B.Tech (C.S.E.) Galgotias College of Engg. and Tech., Greater Noida
E-mail: [1]09.aatif@gmail.com, [2]samie92@gmail.com, [3]sundaramraghuvanshi@gmail.com, [4]trish29aug@gmail.com

Abstract—*The Internet is full of knowledge repositories like Wikipedia and Freebase where every voluntary person has the access to modify or make changes to their data. The made alterations make their way to the web page after they have been reviewed and approved manually. Changes and modifications made by millions of users across the globe require quite an amount of man power for review and assessment. Here we present an implementation of a method called CQUAL once proposed in a paper. This method exploits the historical records of the contributor to predict the contribution quality which comprises of a set of pointers covering multiple fields like professional together with academic and other relevant domains. Alongside, the contribution made by the user goes through a modified TFIDF algorithm to generate a numeric value. The scores obtained from both the algorithms are clubbed together to decide the final credibility of the user and the contribution. This method examines the validity of a contribution immediately after its submission eliminating the task of post submission human assessment notably.*

1. INTRODUCTION

With the inception of Internet, the complete globe and its residents have become dependent on its innumerable features and services. Be it mundane tasks or expert, internet is always there to solve our queries and update us with the best results. From the never ending list of Internet applications, the most significant is accessing information from knowledge repositories. A knowledge repository is a technology for the storage of information by a computer system. Some knowledge repositories are the ones where users can share their information about various domains by uploading data as well as can access the already shared data. Along with it, they can even edit the currently available data on the knowledge repository present for public access.

Uploads from voluntary users in the form of their contributions[1] can sometimes be misleading and provide wrong information. This can arise due to carelessness of the contributor, misunderstanding of the concept or lack of accepted ground truth. Such errors can prove to be disastrous if left unchecked because people accessing these knowledge repositories will be provided with incorrect information that can prove to be highly dangerous. There are several other applications as well such as Google's Knowledge Graph and Bing's Satori which fetch their data from these knowledge repositories. They too will have inaccurate data thus degrading the complete experience.

To eliminate these errors, the uploaded contributions are sent through manual review and assessment. Since they follow post-moderation approach, the contribution goes live immediately and can later be edited or modified by other users. This given period of time between the contributions getting live and the manual review has the maximum possibility of user's access to incorrect information. Also there are quite high chances that even after the manual review; few errors may creep in due to negligence of the editor or malice. Also the manual review requires immense manpower for thousands of changes done on the data of knowledge repositories as well as for the newly uploaded contributions.

Hereby we work on a method called CQUAL[2] which automatically predicts the quality of contributions submitted to a knowledge base as the name itself suggests, 'Contribution QUALity'. This method works over the past knowledge of the contributor which includes his field of expertise, academic record, career prospects, etc. where each has its individual weightage. The overall aim of CQUAL is to decide the credibility of user and his contribution on the basis of his past experience in the respective field in which he presents his contribution.

The complete method works on two individual algorithms. One of them is CQUAL[2] which works on

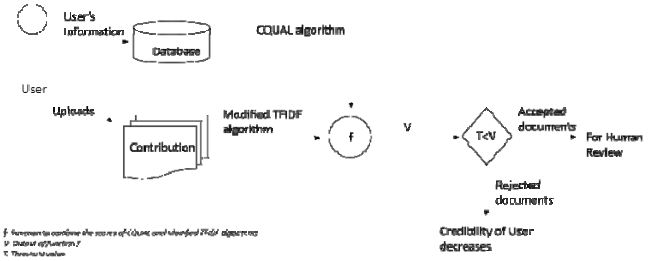

Fig. 1: Working methodology of quality prevision by CQUAL and modified TFIDF algorithms.

user's past domain. Other technique works on the contribution done by the user i.e. the document which he uploads on the knowledge repository or the editing performed by him on the already present data. This employs the modified TFIDF[4-5] algorithm where the relevant keywords are extracted from the document and their frequency is calculated. Further these keywords are mapped by the data dictionary and the list of extracted keywords is again shortlisted. The resulting value is compared to a pre-specified threshold value to either reject or accept the contribution and hence affecting the credibility of the user which is in the form of a numeric value. This further affects the user's profile as it is again selected as a pointer in the form of prior contribution history during the implementation of CQUAL.

It is an automatic moderation technique for verifying user's contribution in real time. This helps in approving large fraction of contributions to be approved instantaneously eliminating the need for human review. Although it is to be noted that this algorithm doesn't provide complete alleviation from manual review. This just reduces the human labor significantly.

2. WORKING

For the quality prevision of contributions for a knowledge base, user's complete history needs to be stored in a database before he makes the contribution on a knowledge repository. One way to accomplish this task is the creation of a form on a web page which asks for every minute detail of the user which can play a vital role in deciding the domain of expertise as well as field of interest of the user. A form can be called good if it provides the user with every possible option itself and where he is always able to find what he needs. For e.g.

while entering the academic record, if a user graduated from say college X, then this should be present beforehand in the list of colleges in the form. It becomes quite necessary because of the fact that every college has a predefined weightage in the database assigned by the administrator. This assignment is based on the fact that how good or bad an institute is compared to other institutes. It can be said that this is a relative marking like the ranking of institutes by a magazine or a media house. This weightage is taken into consideration while calculating the score for CQUAL. The assignment of values is a subjective issue and can vary from one person to another person. For e.g. an institute A may be better than institute B for a group of people whereas others may think of B as a better option than A. Hence, this weightage completely depends on the firm/organization employing this algorithm as it will be the one to decide what weightage should be allotted to an institute/company depending on its requirement and specifications. Also these weightages are time-variant and may change quite frequently which should be noted by the administrator for better functioning and efficiency.

After the complete record of user's past profile has been maintained, the user is allowed to upload its contribution or make changes to the already available data. The contribution goes through modified TFIDF[4-5] algorithm to find the most relevant keywords and their frequency. Along with the most frequently occurring keywords, most frequently co-occurring keywords are also found which gives more efficient results. Further the keywords are mapped by the data dictionary or knowledge base and the final score is generated.

The two values obtained from TFIDF and CQUAL algorithms are clubbed together in a particular proportion which is defined by function f. This proportion once again depends on the organization employing this method but has to follow a certain constraint. This constraint specifies that higher weightage needs to be allotted for the TFIDF's score then the CQUAL's score. For e.g. the weightage of TFIDF and CQUAL can be 70% and 30% respectively. If this proportion is considered and the score of TFIDF and CQUAL after running the respective algorithms are 96 and 54 respectively, then the application of function f gives the value 83.4 (96*0.7+54*0.3). The reason for keeping higher weightage for TFIDF's score and lower weightage for CQUAL's score is because the CQUAL's score is based on user's academics and merits and henceforth we cannot rely completely on the fact that a person with poor merit is not always wrong. That is it is not necessary that a person with a great mind must have scored excellent marks in academia or must have worked in a top ranked firm. Similarly a merit holder may not have an extra brilliant mind. This is the reason why TFIDF's result has always to be given higher priority than CQUAL's result.

The value obtained from the above function has to be finally compared to a threshold value to decide whether the contribution is to be accepted or rejected. This threshold will be a numeric value above which all the documents will be accepted and below which the documents will be rejected. This threshold will have to be decided by the organization itself employing this method. This has to be an optimum value which gives the best results and can be found mainly by making test cases. An approximate value can be decided which has to go through the testing procedure and final value can be calculated from it. At last the accepted documents are sent for manual review enhancing the credibility of the user which is reflected in his profile. This credibility is again a pointer in the form of prior contribution history while implementing the CQUAL algorithm. Similarly the rejected documents which do not satisfy the threshold criteria decrease user's credibility and again reflected in his profile projecting a negative impact for his future contributions.

3. CQUAL

The CQUAL[2] algorithm as already mentioned predicts the quality of contribution submitted to a knowledge base by the user. This is accomplished by creating a database of the user containing his complete history regarding his contributions and expertise of domain. The CQUAL algorithm works on the basis of few pointers which are mainly the parameters on

which the CQUAL score will be calculated. These pointers are picked from the vast history of the user for e.g. his academic performance in school/college, the firm he is employed in, his duration of work experience, etc. Each pointer has a specific set of possible options with each option having a pre-assigned numeric value which acts as the weightage/rank for that option. Once the user fills the form completely the numeric values for each selected option are added together and their average is calculated. This is followed by calculation of its significant value. This gives the final value of the CQUAL algorithm which will be further clubbed together with TFIDF's value to generate the final score deciding whether the document is being accepted or rejected.

The decision of pointers is the task of the organization but we have come up with the following list of most relevant parameters which should always be taken into consideration:

1. Academic record: The academic record of a person although not completely reliable but can be a major deciding factor for CQUAL as it is a pretty good reflection of a person's educational history. It includes user's high school score, college score, subjects, score in individual subjects, etc. For e.g. if we talk about the graduation details of a user, it will consist of his field of graduation, his subjects, his score in individual subjects, etc. This intense detailing gives more precise and efficient results of CQUAL. Greater the information of the user, better the results will be. With graduation, the post-graduation related values can also be stored in the database. Here again like graduation, the major fields will be stream, subjects, scores, etc. Let us talk about a specific case where a person has pursued post-graduation more than once and for sure he should get a higher weightage then the person who pursued post-graduation either once or didn't pursue at all. Hence the input form should be flexible in a way that user should not miss even single of its achievement or academic record. Again if the user is a doctorate he gets a higher weightage which is in turn reflected by numeric value assigned to his 'research' pointer or 'doctorate' pointer. This again is a detailed reflection of user's educational history. Similarly other academic domains can be inculcated.

2. Training/Internship: The internships or trainings pursued by a person in a particular field can be used as parameter to decide the credibility of the user in that field. Related to this, another major deciding factor can be the institute/firm from where the training has been pursued. For e.g. a person in the field of IT industries completing his/her internship from world's best IT firm gets a higher weightage then the one who completes his training from an IT firm localized and restricted in a small city. Along with this the training duration also matters.

3. Certification: Various certification courses are available in respective fields and a person completing a given certification marks his excellence in that field. Similar to the training parameter, a certification completed from an international firm or institute needs to be given a higher weightage then the one completed from a local institute. More the number of certifications, higher the weightage will be.

4. Work Experience: The work experience of a user combines many parameters together. Firstly the duration or time period a user has been employed. Secondly the firm in which he is employed. Thirdly his designation in the firm he is employed. Self-employment needs also to be considered. Overall it can be said that better the firm, better the designation and greater the job duration; higher the weightage.

5. Journals/Surveys published: A person uploading data in a particular field gets a higher weightage if he has published a journal or a research paper previously in the same field. More the published papers higher the weightage.

6. Previous contribution records: A user may have done contributions to the knowledge repositories before out of which few may have been rejected whereas few may have been accepted. This can be one of the deciding factors for his credibility. If majority of the contributions have been rejected it means the user always comes up with a poor knowledge or unreliable information and hence gets a lesser weightage. Similarly if the user's contribution has been accepted majority of the times it can be considered as a good response from the user's end.

7. References: The references provided by a contributor gives an indication of his social life and this information can be gathered in the form of name of the reference, his firm of employment, his designation, etc. The respective person can be contacted and asked for the contributor's information.

4. TFIDF

The second part of this method deals with checking the relevance of the contribution with the respective domain in the knowledge base. For this, the modified TFIDF algorithm[4-5] is used. TFIDF stands for Term Frequency- Inverse Document Frequency. The modified TFIDF algorithm works in two parts. First, it extracts the keywords by removing the

Table 1: Frequency distribution.

Word	data	Mining	knowledge
Frequency	2	2	1

Table 2L Co-occurrence matrix.

	data	Mining	knowledge
data	-	1	0
mining	0	-	0
knowledge	0	0	-

stop-words and also computes the frequency of each keyword in the contribution. The second part of the algorithm tabulates the co-occurrence matrix which results in finding the most frequent co-occurring words in the document. By finding the frequent occurring words and the frequent occurring co-words

we could estimate the relevance[3] of the document with the respective domain thereby reducing the manual work required.

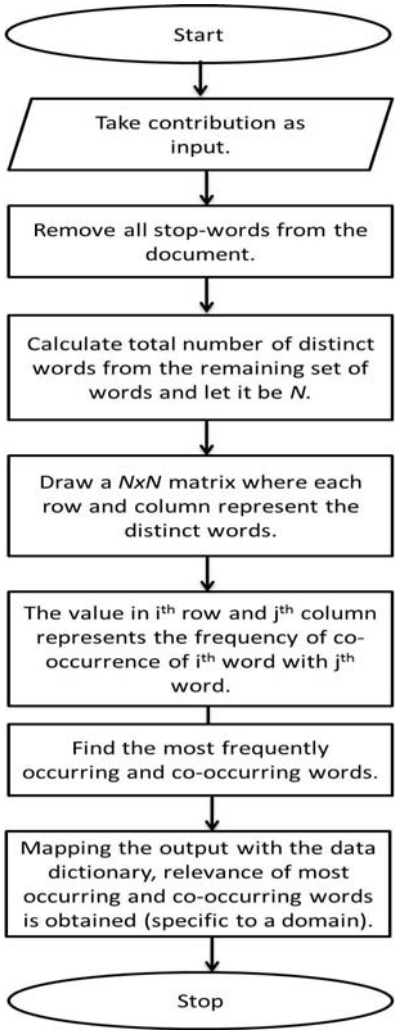

Fig. 2. Flowchart of modified TFIDF algorithm.

5. WORKING OF MODIFIED TFIDF ALGORITHM

Step1: The contribution of the user is processed to remove the stop-words. For e.g. removal of stop-words from the sentence "Data-mining is the mining of knowledge from data." leads us to the set of keywords: 'data', 'mining' and 'knowledge' where the stop-words 'is', 'the', 'of' and 'from' have been removed.

Step2: After removing the stop-words, a list of keywords is obtained and their frequency of occurrence in the document is calculated. As in the above example, the frequency of occurrence of 'data' and 'mining' is 2 each, and of 'knowledge' is 1 (Table 1).

Step3: Next, the frequency of most frequent occurring co-words is tabulated. For this we build a nxn matrix where n is the number of keywords extracted from the document. In this matrix the entry in the ith row and jth column corresponds to the frequency of the number of times the keyword 'j' occurred after keyword 'i' (Table 2).

Step4: The resulting frequent occurring keywords and co-occurring keywords are mapped with the knowledge base of that domain on which the contribution has been made to find the relevance of the document with respect to its domain.

Step5: Finally the common keywords from the contribution and the knowledge base with their respective frequencies are summed together to generate the final score for modified TFIDF algorithm.

6. DISCUSSION

The complete process of CQUAL[2] works on the user's past knowledge which can be provided by him/her by filling up forms on a web page which raises a crucial question of authenticity of the data provided by the user and its validity. One of the methods to eliminate the given problem is to ask the user to send the scanned copies of all the documents confirming the information provided by him in the forms which can be further cross-checked by the administrating firm with the corresponding departments/offices from where the contributor provided the confirmation documents. This undoubtedly seems to be a lengthy task at once and raise the question for the complete CQUAL procedure which aims at eliminating the manual labor (although not fully but partially). For this it has to be noted that the complete verification of user's information needs to be done only once which will generate a score which can be used in future for all the contributions done by him. The score value that is specific to a user sees a change when user modifies any of his information in his profile and only the made modification will need verification, not the complete profile. Hence the CQUAL algorithm although not completely but eliminates the need for manual review quite to an extent.

REFERENCES

[1] B. T. Adler, L. de Alfaro, I. Pye, and V. Raman., "Measuring author contributions to the wikipedia", In 4th Int'l Symposium on Wikis. ACM, 2008.

[2] Chun How Tan, Eugene Agichtein, Panos Ipeirotis, Evgeniy Gabrilovich., "Trust, but Verify: Predicting Contribution Quality for Knowledge Base Construction and Curation", ACM. 2014.

[3] Debajyoti Mukhopadhyay, Pradipta Biswas, Young-Chon Kim., "A Syntactic Classification based Web Page Ranking Algorithm", In MSPT. 2006.

[4] Juan Ramos., "Using TF-IDF to Determine Word Relevance in Document Queries", In ICML. 2003.

[5] Yutaka Matsuo, Mitsuru Ishizuka., "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information", WWW. 2003.