Performance analysis of a keyword search system

Mustafa Abdalrassual Jassim
Department of Civil Engineering , College of Engineering , Muthanna University
Mustafa944@yahoo.com

Abstract

Data mining is the process of discovering patterns in a data set by keyword. Keyword search is the most effective way to discover information in documents. But somewhere, sometimes just searching for a keyword is not enough; with research restricting that keyword has become a necessity. Like in social media abuse of word is increasing. Many systems worked on only detecting an inappropriate word; not on restriction of that word. So here in this paper keyword search method is proposed for social media which not only finds the inappropriate words, but also restrict that word from publishing on the media.

Keywords :- Keyword, Social media, Online social networking, Data mining.

الخلاصة

تعدين البيانات هي عميلة لاكتشاف انماط في مجموعة من البيانات وفقاً للكلمة الرئيسية. البحث عن الكلمة الرئيسية هي الطريق الاكثر فاعلية لاكتشاف المعلومات في الوثائق. ولكن في مكان ما، في بعض الأحيان فقط البحث عن الكلمة الرئيسية ليست كافية، مع البحث في تقييد تلك الكلمة الرئيسية أصبح ضرورة. كما هو الحال في إساءة استخدام وسائل الاعلام الاجتماعية من كلمة آخذت في الازدياد. عملت العديد من الأنظمة على الكشف عن كلمة غير ملائمة فقط؛ وليس على تقييد تلك الكلمة. حتى هنا في ورقة البحث الكلمة المقترحة في طريقة وسائل الاعلام الاجتماعية التي لا يجد فقط الكلمات غير المناسبة، ولكن أيضا تقييد تلك الكلمة من النشر على وسائل الإعلام.

الكلمات المفتاحية: الكلمة ، وسائل الاعلام الاجتماعية، الشبكات الاجتماعية على الانترنت، بيانات التعدين.

1. Introduction

The use of internet is increasing day by day, there are so many peoples who use search engine daily and performs about 4 billion searches. As the searching information is increasing day by day the demand for the keyword search is also increasing rapidly. There are many existing relational keyword search systems. Keyword search systems should return whatever answers they can produce fast. (Kaveri and Naoghare,2015) Data mining is nothing but extraction of data from a large data set. At the time of searching user insert a keyword and get a result associated with that keyword after searching. (Pradeep and Ruhi ,2014) Keyword search depends on applications and retrieval system also differs for that purpose. Requirement of applications change as per its use, also vary according to requirement. (Pradeep and Ruhi ,2014) In this paper the system is being used for social media. As we know that these days Social networks are becoming more and more popular. It is a platform where people can share their thoughts and ideas (Sandeep and Saurabh , 2015). Social media is used for communication purpose. So for communication we must know the name of that person to whom we are chatting. Using login authentication we can get the required name of chatting person from user databases. But main aim of this paper is to restrict the inappropriate word used against any person or religion or community, etc. For that purpose we are also using inappropriate words database in which system not only search the keyword but also restrict that word from publishing. As we know that as popularity of the websites increasing, their vulnerability to be attacked is also increasing(Sandeep and Saurabh ,2015). Message posts by the users can contain some kinds of abusive or offensive contents (Krishna and Narendra , 2015). So it is required to prevent the inappropriate word used against any person or religion or community, etc we are using this system in which each post is checked at run time not after publishing the post.

2. Literature Survey

2.1 Keyword search systems

BANKS: Browsing and Keyword Searching in Relational Databases

Aditya *et.al.,* (Aditya *et.al.,* 2002) has proposed "*BANKS: Browsing and Keyword Searching in Relational Databases",* in which system enables keyword based search on relational databases. "Banks enable the user to accurate information in without any knowledge of the scheme". The user can get information by typing some keywords and then follow the given hyperlinks and get the results.

SPARK: Top-k Keyword Query in Relational Databases

Yi Luo *et.al.,* (Yi Luo and Xuemin , 2007) has proposed, "SPARK: Top-k Keyword Query in Relational Databases" in which they concentrate on the viability and effectiveness of keyword query search. They have proposed a ranking formula using previous information retrieval techniques. The main importance of this technique is that it works on real databases on a large scale, such as Customer Relationship Management (CRM).

Relational Keyword Search System

Pradeep and Ruhi (Pradeep and Ruhi , 2014) has proposed," Relational Keyword Search System" that sets the search word category first and then after the user will select the appropriate word meaning which is going to search. They will statically add the database in their project. It also shows the order of the word and does not require knowledge of database queries.

2.2 Following is the systems used to detect an offensive content.

A System to Detect Inappropriate Messages In Online Social Networks.

ROHAN SHETTY (ROHAN SHETTY *et.al.,*2015) *et.al.,* has proposed *A System To Detect Inappropriate Messages In Online Social Networks* in which they use Support vector machine and naïve Bayes algorithms. Support Vector Machine (svm) determines "an ideal hyper plane and Naive Bayes implements using a probabilistic approach, Naïve Bayes is known to perform" even highly sophisticated classification methods.

Detecting Offensive Language in Social Media to Protect Adolescent Online Safety.

Ying Chen *et.al.,* (Ying , 2012) has proposed, "*Detecting Offensive Language in Social Media to Protect Adolescent Online Safety*" in which they have worked on Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potentiality of users to send out offensive words in social media.

The How, When and Why of Sentiment Analysis.

Mrs. Vijyalaxmi M *et.al.,* (Vijyalaxmi , 2013) has proposed," *The How, When and Why of Sentiment Analysis*" in which they applied Sentiment analysis On a large scale to classify and summarize the review and forecasting.

Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study

Trisha Dowerah Baruah has proposed a (Trisha , 2012) , "*Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study*" in which She has done analysis on the role, the importance and impact of social media as a tool for effective communication. To achieve the effectiveness of social media, the survey method was used to investigate in order to know the increasing importance.

3. Proposed Methodology

In this paper a system is proposed to perform analysis of a keyword search. This system is being used in social media applications. This system is based on in appropriate languages, words used while chatting on the net in social media. When the user types some matter to be posted on the chat application with or without in appropriate words, any person or religion or community, etc in the post, the content will be blocked before being posted. The post will be checked for in appropriate words, if such word in not found, the post is allowed to be posted, otherwise the post is deleted and not published.

In case of a defaulter user will be logged out and his fault count indicator will be incremented by 1. The suspect user will be required to submit an undertaking that he or she will not commit such mistake in future. Three chances will be given to the user in similar way and will be required to submit the certificate on the mischief and fourth time his account will be blocked permanently. His login credentials will be saved and he or she will not be allowed to be a member of the chat group.

- For implementation we are using association rule mining
- Classification Based On Association Rule Mining

In associative classification, the focus is to produce association rules that have only a particular Attribute in the consequent. These association rules produced are called Class Association Rules (CARs).

Table 1: Training Data

Row#	AT1	AT2	Class
1	Z1	W1	P1
2	Z1	W2	P2
3	Z1	W1	P2
4	Z1	W2	P1
5	Z2	W1	P2
6	Z2	W1	P1
7	Z2	W3	P2
8	Z1	W3	P1
9	Z2	W4	P1
10	Z3	W1	P1

- Let T be the training data set with m attributes AT1, AT2, … , ATm and |T| rows. "Let P be a list of class labels. An item is defined by the association of an attribute and its value" (ATi, ai), or "a combination of between 1 and m different attributes values. A rule r for classification is represented in the form": $(AT_{i1} = x_{i1}) \wedge (AT_{i2} = x_{i2}) \wedge \ldots \wedge (AT_{in} = x_{in}) \rightarrow p_{i1}$ where the antecedent of the rule is an item and the consequent is a class.
- This algorithm uses a set of relevant rules to make a prediction decision by evaluating the correlation between them

3.1 Block Diagram

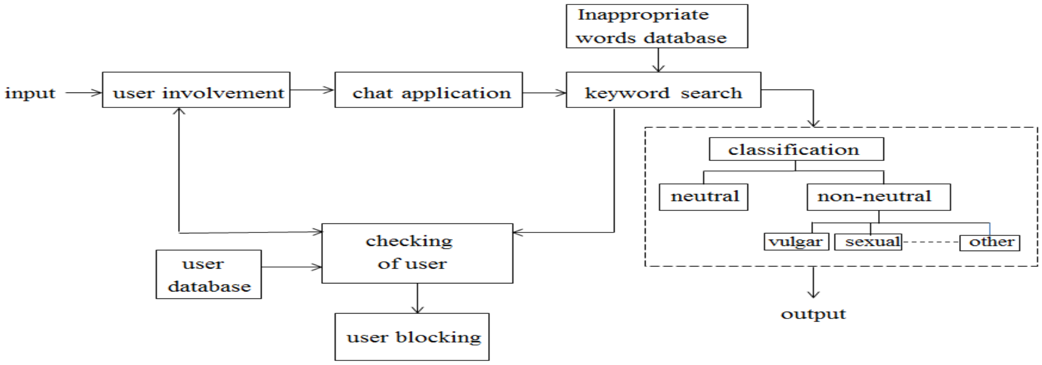


Fig 1: Block Diagram

Modules

User involvement, User checking, Chat application, Classification of Content, Checking of statements, User blocking

1. **User involvement:-**In User involvement login of an authorized person is done by using user name and password received after properly enrolling into the chat room.
2. **User checking module:-**On login user checking module checks the database of users. If the defaulter variable count for that particular person is three then user is not allowed to chat, and for new user or the users who has a count less than three is allowed to chat.
3. **Chat application:-**To use the chat application first user has to login with their user name and password. By checking the authentication and the database of that particular user, chat application will be open for them in which user is able to send data, receive data.
4. **Classification of Content:-**In content filter, it first takes input statement written by the user on chat room. Then make a tree with the words of post. Tree search is efficient searching method and it required less time.
5. **Checking of the statement:-**In this module remaining words are checked with database of inappropriate words if there is any inappropriate word found, then count for that person will be incremented by one and after formalities a chance is allotted for that person to use the application. This chance will be incrementing his or her fault count by one and his or her chance to do mistake will be decreased by one.
6. **User blocking:-**In this module person is blocked when the frequency of using inappropriate word by that person become~~s~~ three.

3.2 Algorithm

Step 1: User login

Step 2: Allocating chat environment to the authorized user.

Step 3: User input into chat application

Step 4: On clicking post button statement will be transferred to the filter.

Step 5: Filter will make a tree of words for smooth check-in.

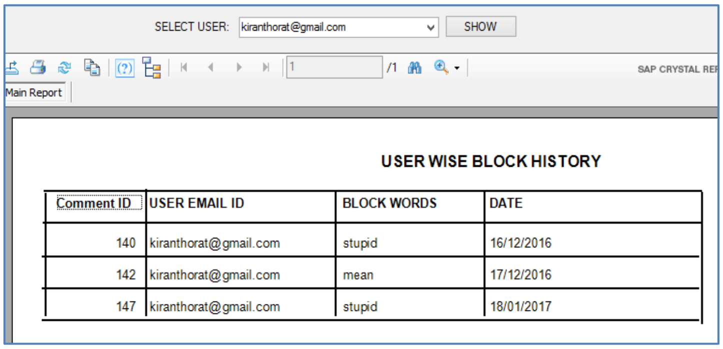
Step 6: Transfer of this tree to statement checking module.

Step 7: If content doesn’t match with database, statement is allowed to move out; else if content matched them

a) User data are checked from the user database.
b) Necessary action will be taken.

4. Proposed Prototype

1. **User involvement:-**User involvement contains all the details of users like user name, date of birth, Email address, city country, password etc. First user has to register here, after successfully registration he can enter into chat room with mail id and password.
2. **User checking:-**In user checking module on login it checks the database of users. If the defaulter variable count for that particular person is three then user is not allowed to chat, and for the new user or the users who has a count less than three is allowed to chat.
3. **Chat application:-**To use the chat application first user has to login with their user name and password. By checking the authentication and the database of that particular user, chat application will be open for them in which user is able to send data, receive data.
4. **Keyword search:-**In this module each and every word is checked with database of inappropriate words if there is any inappropriate word found then defaulter count for that person will be incremented by one and after formalities a chance is allotted for that person to use the application. This chance will be incrementing his or her fault count by one and his or her chance to do mistake will be decrease by one.

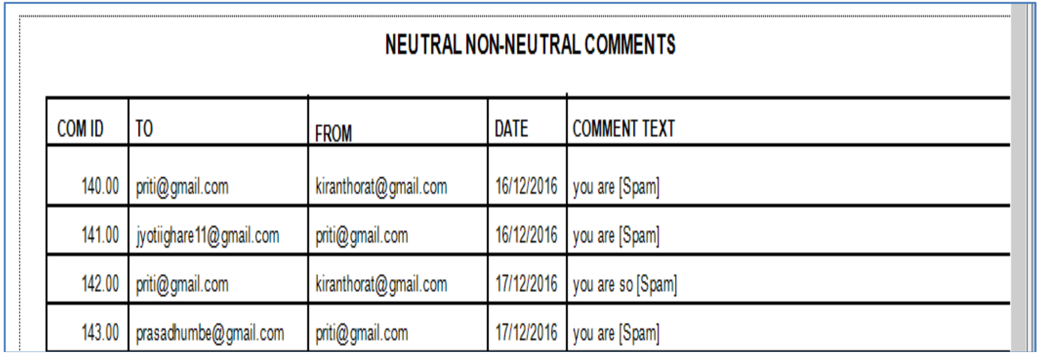


SELECT USER: kiranthorat@gmail.com

USER WISE BLOCK HISTORY

Comment ID	USER EMAIL ID	BLOCK WORDS	DATE
140	kiranthorat@gmail.com	stupid	16/12/2016
142	kiranthorat@gmail.com	mean	17/12/2016
147	kiranthorat@gmail.com	stupid	18/01/2017

Fig 2: Data Analysis snapshot



NEUTRAL NON-NEUTRAL COMMENTS

COM ID	TO	FROM	DATE	COMMENT TEXT
140.00	priti@gmail.com	kiranthorat@gmail.com	16/12/2016	you are [Spam]
141.00	jyotiighare11@gmail.com	priti@gmail.com	16/12/2016	you are [Spam]
142.00	priti@gmail.com	kiranthorat@gmail.com	17/12/2016	you are so [Spam]
143.00	prasadhumbe@gmail.com	priti@gmail.com	17/12/2016	you are [Spam]

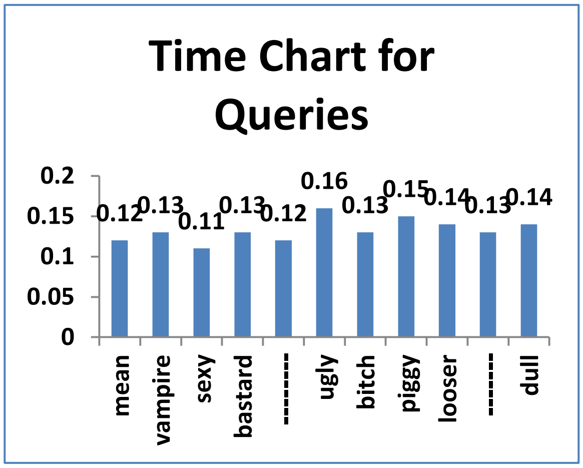
Fig 3: Summary of Query fired

5. **User blocking:-**In this module person is blocked when the frequency of using inappropriate word by that person become three.

5. Data Analysis

Table 2: Processing Data & Chart

Comment	Abusive word	Execution time
you are so dumb	dumb	0.13
Ravi is a thief	thief	0.15
how funny you are looking in pic	funny	0.13
Happy birthday	-------	o.12
How mean she is	mean	0.12
She is a vampire	vampire	0.13
priyankalooks sexy in a film	sexy	0.11
You bastard, I will kill you	bastard	0.13
Looking beautifull	--------	0.12
you ugly woman	ugly	0.16
You bitch	bitch	0.13
I called him piggy	piggy	0.15
Hey looser	looser	0.14
Happy diwali	-------	0.13
deepika is looking dull in movie	dull	0.14



On the contents of table 2, we see the chart that shows tics required for searching and blocking of any user who has written in his post any kind of word which is banned or is a bad word. The key word search using this prototype model gives results at par with the latest results as per the Joel Coffman and Alfred C. Weaver (Joel and Alfred , 2014). In their research they discussed the keyword search algorithm which gave good results in terms of time complexity. Our results at Intel(R) Core (TM)2 Duo CPU T5550 @ 1.83GHz, with 2GB internal RAM machine on 32bit Windows OS environment and Dot Net platform, are more positive. This can be further enhancesd by replacing variable parameters. This can be done as a future work.

6. Conclusion

In existing systems, we have seen that all modules have worked on searching the keyword of databases with an additional restriction of keyword being used in the chat on social media. In this paper a keyword is searched; condition checked; if not favourable then restricted to be posted. Hence this system can be termed as unique. This system can be further improved by using self-learning approach or neural network to understand a “Taunt” or a double meaning word. Even punctuations in a sentence also plays a greater role, which is also needed to be understood.

References

Aditya , 2002 ,” BANKS: Browsing and Keyword Searching in Relational Databases”, Proceedings of the 28th VLDB Conference, Hong Kong, China

Joel Coffman and Alfred C. Weaver, 2014, “An Empirical Performance Evaluation of Relational Keyword Search Systems”, IEEE Transaction on Knowledge and Data Engineering,Vol 26, Issue 1

Kaveri A. Dighe and M. M. Naoghare, 2015,”Evaluating Performance of Keyword Search Systems”,International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, Volume 4 Issue 10

Krishna B. Kansara and Narendra M. Shekokar,2015,” A Framework for Cyberbullying Detection in Social Network”, International Journal of Current Engineering and

Technology, E-ISSN 2277 – 4106, P-ISSN 2347 – 5161, Accepted 25 Feb 2015, Available online 28 Feb 2015, Vol.5, No.1

Mrs. Vijyalaxmi M, 2013," The How, When and Why of Sentiment Analysis", Sangeeta Oswal et.al., Int.J.Computer Technology & Applications,Vol 4 (4),660-665, ISSN:2229-6093, IJCTA

Pradeep M. Ghige and Prof. Ruhi R. Kabra, 2014," Relational Keyword Search System", International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014,ISSN 2091-2730.

ROHAN SHETTY, 2015," A System To Detect Inappropriate Messages In Online Social Networks",Proceedings of 18th IRF International Conference, 11th January 2015, Pune, India, ISBN: 978-93-84209-82-7

Sandeep Kumar Rawat and Assistant Prof. Saurabh Sharma, 2015," A Review on Spam Classification of Twitter Data Using Text Mining and Content Filtering" International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 5, Issue 6

Trisha Dowerah Baruah, 2012," Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study", International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012 ,ISSN 2250-3153

Yi Luo, Xuemin Lin, 2007, SPARK: Top-k Keyword Query in Relational Databases", SIGMOD'07 , Chicago,China

Ying Chen, 2012," Detecting Offensive Language in Social Media to Protect Adolescent Online Safety", Published in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (Social Com), (September 2012), pp. 71-80.