ACSIJ Advances in Computer Science: an International Journal, Vol. 5, Issue 1, No.19 , January 2016
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Ranking user's comments by use of proposed weighting method

Zahra Hariri[1]

[1] **Student in Faculty of Graduate Studies, Safahan Institute of Higher Education, Isfahan, Iran, P.O.Box 817474319**
Zahara.hariri2003@gmail.com

Abstract

The main challenges which are posed in opinion mining is information retrieval of large volumes of ideas and categorize and classify them for use in related fields. The ranking can help the users to make better choices and manufacturers in order to help improve the quality. As one of the pre-processing techniques in the field of classification, weighting methods have a crucial role in ranking ideas and comments. So, we decided to offer a new weighting method to improve some other similar methods, especially Dirichlet weighting method. In this paper, the proposed method will be described in detail, and the comparison with the three weighting methods: Dirichlet, Pivoted and Okapi also described. The proposed weighting method has higher accuracy and efficiency in comparison to similar methods. In the following, user comments of online newspapers are ranked and classified by use of proposed method. The purpose is to provide more efficient and more accurate weighting method, therefor the results of ranking will be more reliable and acceptable to users.

Keywords: *Opinion mining, Information retrieval, ranking comments, weighting methods, weighting methods constraints.*

1. Introduction

Word Wide Web can be considered as a repository of ideas from users. The challenge that the manufacturers and web administrators are faced by is to analyze and organize their ideas.

Analysis of emotions in online publications is a way of organizing user's ideas, which requires weighting of the words in comments. The weighting methods include genetic algorithms, artificial neural networks, regression equations, TF-IDF, Pivoted, Okapi, Dirichlet. In this article we propose a new weighting method which is improved of Dirichlet weighting method, also it satisfy all 7 constraints.

The paper is organized as follows. Section 2 state some weighting methods and, also proposed weighting method. All the constraints are checked for the proposed method in section 3. In section 4, accuracy and performance of the proposed method is compared with methods such as: Pivoted, Okapi and Dirichlet. Section 5 is about implementation of ranking comments by use of proposed weighting method. Dataset is discussed in section 6. Conclusion and future works are expressed in section 7. Finally, section 8 states references.

This document is set in 10-point Times New Roman. If absolutely necessary, we suggest the use of condensed line spacing rather than smaller point sizes. Some technical formatting software print mathematical formulas in italic type, with subscripts and superscripts in a slightly smaller font size. This is acceptable.

2. Weighting methods

There are a lot of weighting methods that are used. But there are different in 2 aspects: satisfying constraints and the value of parameters like efficiency and accuracy. In the following some weighting methods are mentioned.

2.1 Pivoted method

Vector space model is displayed as a vector of words. Documents are ranked based on the similarity between query and document vector. Pivoted retrieval method is one of the best retrieval formula which is expressed in equation 1 as bellow [1].

$$S(Q,D) = \sum_{t\in Q\cap D} \frac{1+\ln(1+\ln(c(t,D)))}{(1-S)+S\frac{|D|}{awdl}} . c(t,Q). ln \frac{N+1}{df(t)} \quad (1)$$

Where S is retrieval parameter, c(t,D) is The number of repetitions of word t in document D. |D| is the length of document D. c(t,Q) is The number of repetitions of word t in query Q. N is the number of documents and df(t) is the number of documents including word t.

2.2 Okapi method

This formula is an effective retrieval formula that uses classical probabilistic model. It is expressed in equation 2[1].

$$S(Q,D) = \sum_{t\in Q\cap D} ln\frac{N-df(t)+0.5}{df(t)+0.5} \times \frac{(k_1+1)\times c(t,D)}{k_1((1-b)+b\frac{|D|}{awdl})+c(t,D)} \times \frac{(k_3+1)\times c(t,Q)}{k_3+c(t,Q)} \quad (2)$$

K_1 is between 1 and 2 , b is equal to 0.75 , K_3 is between 0 and 1000. df(t) is the number of documents including word t, c(t,D) is The number of repetitions of word t in document D and awdl is the average of document's length.

2.3 Dirichlet prioir method

This method is one of the best methods of language modeling approach which working based on similarity between query and document. It is expressed in equation 3[1].

$$s(Q,D) = \sum_{t \in Q \cap D} c(t,Q).\ln(1 + \frac{c(t,D)}{\mu.p(t|C)} + |Q|.ln\frac{\mu}{|D|+\mu} \quad (3)$$

Where μ is retrieval parameter, c(t,D) is The number of repetitions of word t in document D. |D| is the length of document D and |Q| is the length of query Q. $p(t|C)$ is possibility of existence of word t in the collection.

2.4 Proposed weighting method

In this study a new weighting method is proposed in equation 4 which aim is to satisfy all the constraints and, also improve accuracy and efficiency of previous methods such as Pivoted, Okapi and Dirichlet.

$$S(Q,D) = \sum_{t \in Q \cap D} (C(t,Q).C(t,D).\ln\left(1 + \frac{C(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D|} \qquad (4)$$

Where μ is retrieval parameter, c(t,D) is The number of repetitions of word t in document D. |D| is the length of document D and |Q| is the length of query Q. c(t,Q) is The number of repetitions of word t in query Q.
Dirichlet method don't satisfy the LNC2 constraint but proposed method satisfy all the weighting method constraints which will be explain in the next section.

3.CHEKING WEIGHTING METHOD CONSTRAINTS FOR PROPOSED METHOD

There are 7 constrains which are good to be satisfied by weighting methods. In the following all 7 constraints are checked for proposed method.[1,2]

3.1. TFC1

In equation 5, it is shown that proposed method satisfies TFC1.

$$s(Q,D_2) = c(t,Q).c(t,D_2).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_2|}$$

$$s(Q,D_1) = c(t,Q).c(t,D_1).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_1|}$$

C(t,D_1) > C(t,D_2) then s(Q,D_1) > s(Q,D_2). (5)

3.2. TFC2

In equation 6, it is shown that proposed method satisfies TFC2.

$$\begin{aligned}
s(Q,D_2) &= c(t,Q).c(t,D_2).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) \\
&\quad + \frac{|Q|}{|D_2|}s(Q,D_3) \\
&= c(t,Q).c(t,D_3).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) \\
&\quad + \frac{|Q|}{|D_3|}s(Q,D_1) \\
&= c(t,Q).c(t,D_1).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) \\
&\quad + \frac{|Q|}{|D_1|}c(q,D_2) = 1 + c(q,D_1) \\
&\rightarrow \ c(q,D_2) > c(q,D_1) \qquad 1c(q,D_3) \\
&= 1 + c(q,D_2) \ \rightarrow \ c(q,D_3) \\
&> c(q,D_2)
\end{aligned}$$

$$c(q,D_1) < c(q,D_2) < c(q,D_3) \ \rightarrow \ 2c(q,D_1) < c(q,D_2) < c(q,D_3)$$

$$s(Q,D_2) - s(Q,D_1) > s(Q,D_3) - s(Q,D_2)$$

$$\left[c(t,Q).c(t,D_2).\ln\left(1 + \frac{c(t,D_2)}{\mu.df(t)}\right) + \frac{|Q|}{|D_2|}\right] - \left[c(t,Q).c(t,D_1).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_1|}\right] > \left[c(t,Q).c(t,D_3).\ln\left(1 + \frac{c(t,D_3)}{\mu.df(t)}\right) + \frac{|Q|}{|D_3|}\right] - \left[c(t,Q).c(t,D_2).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_2|}\right]$$

$$c(t,Q).c(t,D_2) - c(t,Q).c(t,D_1) > c(t,Q).c(t,D_3) - c(t,Q).c(t,D_2)$$

$$\rightarrow \ c(t,D_2) - c(t,D_1) > c(t,D_3) - c(t,D_2) \ \rightarrow \ 2c(t,D_2) > c(t,D_3) + c(t,D_1) \qquad (6)$$

3.3. TFC3

In equation 7, it is shown that proposed method satisfies TFC3.

$$s(Q,D_1) = c(t,Q).c(t,D_1).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_1|}s(Q,D_2) =$$

$$c(t,Q).c(t,D_2).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_2|}s(Q,D_1) > s(Q,D_2) \ \rightarrow$$

$$s(Q,D_1) = c(t,Q).c(t,D_1).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_1|} >$$

$$s(Q,D_2) = c(t,Q).c(t,D_2).\ln\left(1 + \frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_2|}td(q_1) =$$

$$td(q_2)c(t,Q).c(t,D_1) > c(t,Q).c(t,D_2) \ \rightarrow \ c(t,D_1) > c(t,D_2) \qquad (7)$$

3.4. TD

In equation 8, it is shown that proposed method satisfies TDC.

$$\begin{cases} t \notin Q \rightarrow c(t,D_2) = c(t,D_1)+1 \rightarrow s(Q,D_1) > s(Q,D_2) \\ t \in Q \rightarrow c(t,D_2) = c(t,D_1) \rightarrow s(Q,D_1) = s(Q,D_2) \end{cases}$$
$$\rightarrow s(Q,D_1) \geq s(Q,D_2) s(Q,D_1)$$
$$= c(t,Q).c(t,D_1).\ln\left(1+\frac{c(t,D)}{\mu.df(t)}\right)+\frac{|Q|}{|D_1|}s(Q,D_2)$$
$$= c(t,Q).c(t,D_2).\ln\left(1+\frac{c(t,D)}{\mu.df(t)}\right)+\frac{|Q|}{|D_2|}$$
$$c(t,D_2) = c(t,D_1) \text{ then } s(Q,D_1) = s(Q,D_2). \quad (8)$$

3.5. LNC1

In equation 9, it is shown that proposed method satisfies LNC1.

$$\begin{cases} t \notin Q \rightarrow c(t,D_2) = c(t,D_1)+1 \rightarrow s(Q,D_1) > s(Q,D_2) \\ t \in Q \rightarrow c(t,D_2) = c(t,D_1) \rightarrow s(Q,D_1) = s(Q,D_2) \end{cases} \rightarrow$$
$$s(Q,D_1) \geq s(Q,D_2) s(Q,D_1) = c(t,Q).c(t,D_1).\ln\left(1+\frac{c(t,D)}{\mu.df(t)}\right)+\frac{|Q|}{|D_1|}s(Q,D_2) = c(t,Q).c(t,D_2).\ln\left(1+\frac{c(t,D)}{\mu.df(t)}\right) + \frac{|Q|}{|D_2|} \quad (9)$$

3.6. LNC2

In equation 10, it is shown that proposed method satisfies LNC2.

$$s(Q,D_2) = c(t,Q).c(t,D_2).\ln\left(1+\frac{c(t,D)}{\mu.df(t)}\right)+$$
$$\frac{|Q|}{|D_2|}s(Q,D_1) = c(t,Q).c(t,D_1).\ln\left(1+\frac{c(t,D)}{\mu.df(t)}\right)+$$
$$\frac{|Q|}{|D_1|}c(w,D_1) = K.c(w,D_2) \rightarrow c(w,D_1) > c(w,D_2) \quad 1 \quad (10)$$

Due to equation 1 of 10, $s(Q,D_1)>s(Q,D_2)$, $\frac{|Q|}{|D|}$ will not grow as much as $c(w,D)$, then $s(Q,D_1) \geq s(Q,D_2)$.

3.7. TF-LNC

In equation 11, it is shown that proposed method satisfies TF-LNC.

$c(q,D_1) > c(q,D_2)$ then $c(q,D_1) - c(q,D_2)>0$ and, also $|D_1| = |D_2| + c(q,D_1) - c(q,D_2)$ then $|D_1| > |D_2|$ (11)

Table 1: Weighting method constraints

Formula	*TFC1*	*TFC2*	*TFC3*	*TDC*	*LNC1*	*LNC2*	*TF-LNC*
Pivoted	Y	Y	Y	Y	Y	C	C
Dirichlet	Y	Y	Y	Y	Y	C	Y
BM25	C	C	C	Y	C	C	C
PL2	C	C	C	C	C	C	C
Proposed Method	Y	Y	Y	Y	Y	Y	Y

4. Comparison of Accuracy and Efficiency

To compare the proposed method with three methods which were mentioned before, we need confusion matrix that are explained in next parts. Also we need TN, TP, FN, FP parameters owing to calculating accuracy and efficiency.[3,4]

TN: Are correct but have been misdiagnosed by machine.
TP: Are correct and have been diagnosed correctly by machine.
FN: Are false and have been misdiagnosed by machine.
FP: Are false but have been diagnosed correctly by machine.

Accuracy and efficiency can be calculated by use of equations 12, 13 and 14.[5,6,7]

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \quad (12)$$

$$efficiency = \frac{2*Recall*Accuracy}{Recall+Accuracy} \quad (13)$$

$$Recall = \frac{TP}{TP+FP} \quad (14)$$

Calculating parameters which are necessary from confusion matrix, is shown in tables 2, 3, 4 and 5.

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 25 & 234 & 96 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Okapi confusion matrix

Table 2: Okapi parameters from confusion matrix

	First row	*Second row*	*Third row*	*Forth row*
TN	234	0	234	234
TP	0	234	0	0
FN	0	124	0	0
FP	25	0	96	3

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 275 & 0 & 0 \\ 0 & 0 & 3 & 17 \end{bmatrix}$$

Dirichlet confusion matrix

Table 3: Dirichlet parameters from confusion matrix

	First row	*Second row*	*Third row*	*Forth row*
TN	27	17	27	17
TP	0	10	0	3
FN	1	0	275	0
FP	0	275	1	0

$$\begin{bmatrix} 391 & 1 & 11 & 5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Pivoted confusion matrix

Table 4: Pivoted parameters from confusion matrix

	First row	*Second row*	*Third row*	*Forth row*
TN	0	391	391	391
TP	391	0	0	0
FN	17	0	0	0
FP	0	1	11	5

$$\begin{bmatrix} 72 & 0 & 0 & 0 \\ 0 & 173 & 0 & 0 \\ 0 & 0 & 37 & 0 \\ 0 & 0 & 13 & 11 \end{bmatrix}$$

Proposed method confusion matrix

Table 5: Proposed method parameters from confusion matrix

	First row	*Second row*	*Third row*	*Forth row*
TN	221	120	256	282
TP	72	173	37	11
FN	0	0	0	13
FP	0	0	13	0

For computing confusion matrix, we need a matrix which its rows are documents and columns are words that remain after preprocessing. Equation 15 shows an example.

$$\begin{bmatrix} -1.6708 & -1.6708 & -1.6708 & -1.6707 \\ -1.6708 & -1.6707 & -1.6708 & -1.6708 \\ -1.6708 & -1.6708 & -1.6708 & -1.6707 \\ -1.6708 & -1.6708 & -1.6708 & -1.6707 \end{bmatrix} \quad (15)$$

Then this matrix will be an input for matlab for computing confusion matrix. After that accuracy and efficiency are computable. Table 6 shows the results of accuracy and efficiency comparison.

Table 6: Comparison of accuracy and efficiency

Formula	Confusion Matrix	Accuracy	Efficiency
Okapi	$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 25 & 234 & 96 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	80%	31%
Pivoted	$\begin{bmatrix} 391 & 1 & 11 & 5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	97%	36%
Dirichlet	$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 275 & 0 & 0 \\ 0 & 0 & 3 & 17 \end{bmatrix}$	50%	47%
Proposed Method	$\begin{bmatrix} 72 & 0 & 0 & 0 \\ 0 & 173 & 0 & 0 \\ 0 & 0 & 37 & 0 \\ 0 & 0 & 13 & 11 \end{bmatrix}$	97%	90%

A code that give us an input matrix for computing confusion matrix has some steps as below:

1. A query is written by user.
2. Eliminating stop words.
3. Allocating weight to words by use of one of weighting methods.
4. Creating output matrix(which is input in matlab to calculate confusion matrix)

In figure 1 a schema of an output matrix based on proposed weighting method is shown.

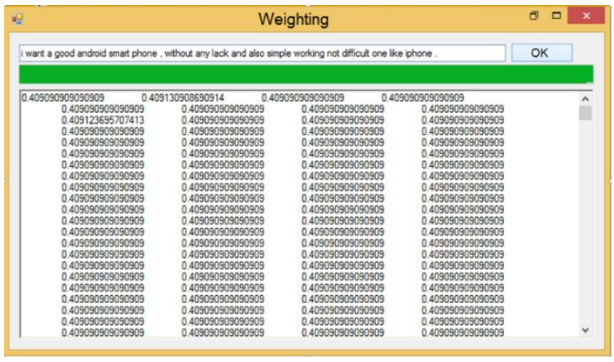

Fig. 1 weighting matrix.

To calculate confusion matrix in matlab some steps has been taken:

1. Test set is created.
2. A random order is created.
3. Sorting input matrix and test matrix based on random order which was created before.
4. Learning set is created.
5. Sample set is classified based on test and learning sets.
6. Confusion matrix is created.

Figure 2 shows a confusion matrix based on proposed method in matlab.

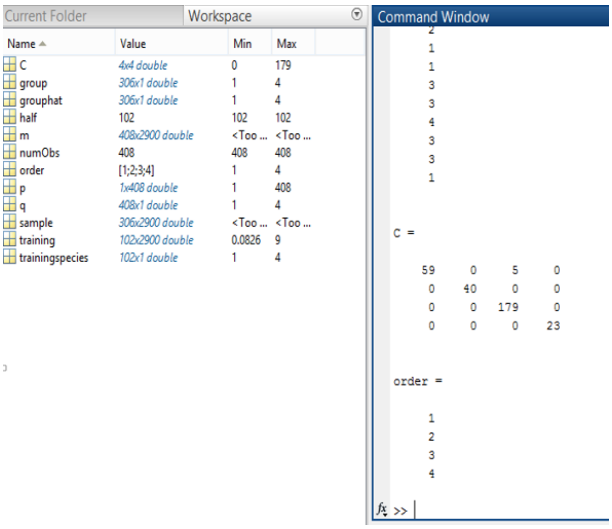

Figure. 2 Confusion matrix in Matlab.

5. Implementation

Text mining and sentiment analysis such as analyzing user's comments can be implemented by using c#.net programming framework. In figure 3 shows a summarized flowchart of an implemented code for ranking user's comments based on proposed method.

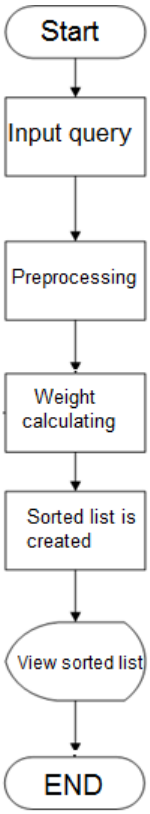

Figure. 3 Implemented code flowchart.

Implemented code is consist of 2 parts, dataset preparation and ranking comments. As is shown in figure 3, first of all, user insert a query in the weighting part and query is sent to preprocessing part. After that, words are weighted based on proposed method. Then a sorted list is created by use of cosine similarity.

Overally, all the steps that has been taken due to ranking comments are as follow:

1. A query is written by user.
2. Eliminating stop words.
3. Price and property of a product is inserted by user.
4. Words are weighted based on proposed method.
5. Documents are ranked based on cosine similarity.
6. Ranked comments are shown.

Implemented code is in c#.net framework for ranking user's comments.

First user insert a query, price and property after that the ok key is selected all of the calculations and computations are done. Then, cosine similarity is calculated by use of equation 16.[8,9]

$$\text{Similarity}=\frac{\vec{d}j \cdot \vec{q}}{|\vec{d}j|.|\vec{q}\ |} = \frac{\sum_{i=1}^{t}(w_{ij}.w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 . \sum_{i=1}^{t} w_{iq}^2}} \quad (16)$$

Where $\vec{d}j$ is document vector, $\vec{q}$ is query vector, $|\vec{d}j|$ is size of document vector, $|\vec{q}\ |$ is size of query vector. w_{ij} is a weight of word I in document j and w_{iq} is a weight of word I in query q.

Figure 4 is an example of Proposed product to user based on below query and ranked comments by use of proposed weighting method.

Query: I want a good android smart phone, without any lack and also simple working not difficult one like iPhone.

Samsung Galaxy S6 Dual-SIM
Samsung Galaxy Note Edge
Samsung Galaxy S4 SGH
Huawei Ascend Y530

Figure. 4 Sample output.

6. **Dataset**

We couldn't find profitable dataset, so we collect a dataset from Amazon.com in the period of times about 2 months. This dataset is about cellphones by 2 property (price and operating system type). Figure 5 shows dataset program's flowchart.

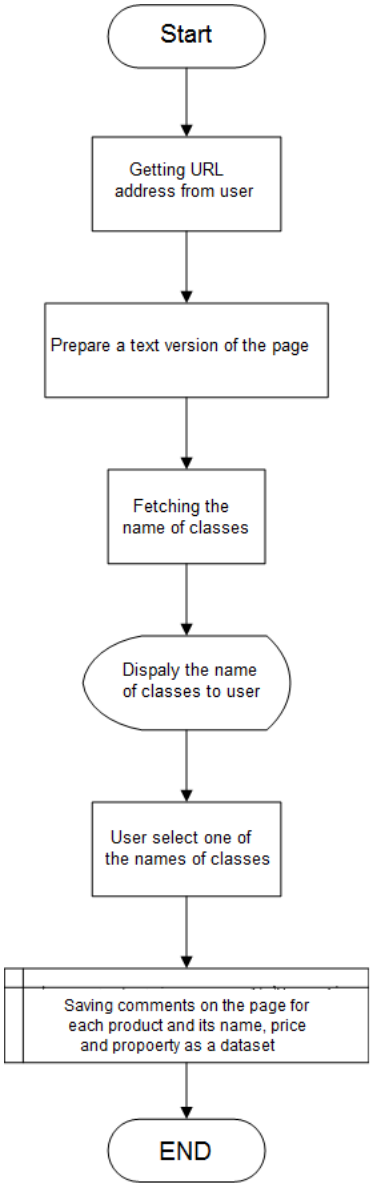

Figure. 5 Gathering Dataset flowchart.

As is shown in figures 6 and 7, all the steps are such as figure 6.

Figure. 6 Gathering Dataset program (steps 1 and 2).

Figure. 7 Gathering Dataset program (step 3)

7. Conclusion

One of the crucial usage of weighting methods is in text mining and information retrieval. Nowadays, ranking user's comments plays a vital role owing to its important help to users for selecting the best product and also help the producer to know best about their products in user's point of view.

Using one of the weighting methods is so important due to ranking comments. Weighting methods are comparable in 2 aspects. Proposed method in this research can satisfy these 2 aspect as well. It can satisfy all the 7 constraints and also it has better accuracy and efficiency in comparison to similar methods such as Okapi, Pivoted and Dirichlet. Another advantage of this research is its recommendation to user about a product he needs.

Figure 8 shows a comparison between proposed method and similar ones.

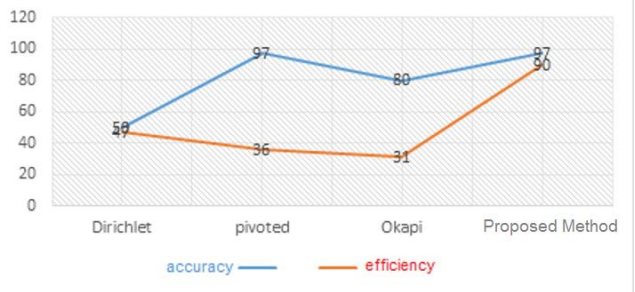

Figure. 8 Comparison of accuracy and efficiency.

Generally, this research's benefits are:

1. User can specify the category of the product.
2. User can determine 2 important property about product.
3. Others comments can be used in field of each product.
4. Weighting methods which was proposed has a better accuracy and efficiency in comparison to others.

Although obtained results show a good performance of the proposed method, we can't claim that it is the best method. The aim of this study was to use the results to provide useful suggestions to the user, but the results can be used for other purposes, too.

Some future works are:

- Improve executive order to enhance the speed of ranking.
- Use proposed method for showing results to the owners of online communities.
- Integrate database issues and user personal profile's information, in order to omitting the stage of sending gathered information from user.

Acknowledgments

"This article is fetched from the thesis in Faculty of Graduate Studies, Safahan Institute of Higher Education."

References

[1] HUI FANG, TAO TAO, CHENGXIANG ZHAI, "Diagnosti c Evaluation of Information Retrieval Models",ACM Transaction on Information Systems, Vol 29, No 2, Article 7,April 2011.

[2] Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida," Ranking news articles based on popularity Prediction", International Conference on Advances in Social Networks Analysis and Mining, IEEE/ACM ,2012, pp 106 – 110, ISBN/ISSN: 978-1-4673-2497-7 .

[3] Chiao-Fang Hsu, Elham Khabiri, James Caverlee," Ranking Comments on the Social Web",Proceedings of the 2009 International Conference on Computational Science and Engineering , Vol 04, ISBN: 978-0-7695-3823-5 , pp 90-97 .

[4] Robertson, A.M. and Willett, P., "An Upperbound to the Performance of Ranked-Output Searching: Optimal Weighting of Query Terms Using a Genetic Algorithm", Journal of Documentation, Vol. 52, pp. 405–420, 1996.

[5] Ghose, A. and P. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In Proceedings of the International Conference on Electronic Commerce, 2007.

[6] Giorgos Giannopoulos, Ingmar Weber, Alejandro Jaimes, Timos Sellis" Diversifying User Comments on News Articles",Lecture Notes in Computer Science Vol 7651, 2012, pp 100-113.

[7] Stefan Siersdorfer, Sergiu Chelaru, Jose San Pedro," How Useful are Your Comments?- Analyzing and Predicting YouTube Comments and Comment Ratings",WWW '10 Proceedings of the 19th international conference on World wide web ,2010,ISBN: 978-1-60558-799-8 , pp 891-900 .

[8] Pooja Kherwa, Arjit Sachdeva, Dhruv Mahajan Nishtha Pande, Prashast Kumar, "An approach towards comprehensive sentimental data analysis and opinion mining", Advance Computing Conference (IACC), 2014 IEEE International, 21-22 Feb. 2014,ISBN: 978-1-4799-2571-1,pp:606-612.

[9] Liu, Bing, "Sentiment analysis: A multi-faceted problem". IEEE Intelligent Systems, 2010.25,3: 76-80.

First Author bachelor degree was achieved in 2012 from TABARESTAN Chalus , Iran in computer software engineering , Student of master degree in SAFAHAN , Isfahan , Iran. Working on Text mining, Sentiment analysis, Public Opinion Research.