World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:2, No:1, 2008

Genetic Mining: Using Genetic Algorithm for Topic based on Concept Distribution

S. M. Khalessizadeh, R. Zaefarian, S.H. Nasseri, and E. Ardil

***Abstract*—**Today, Genetic Algorithm has been used to solve wide range of optimization problems. Some researches conduct on applying Genetic Algorithm to text classification, summarization and information retrieval system in text mining process. This researches show a better performance due to the nature of Genetic Algorithm. In this paper a new algorithm for using Genetic Algorithm in concept weighting and topic identification, based on concept standard deviation will be explored.

***Keywords*—**Genetic Algorithm, Text Mining, Term Weighting, Concept Extraction, Concept Distribution.

Open Science Index, Mathematical and Computational Sciences Vol:2, No:1, 2008 publications.waset.org/10232/pdf

I. Introduction

RAPID growth of available data in digital format increase need for methods to analyze them [13]. So research on some topics such as text classification, information retrieval and automatic text summarization became an important field [8]. Researchers in *Knowledge Discovery in Databases* (KDD) have provided new tools for analyzing and accessing data in databases [7], [12]. Some of them is based on term frequency and are used in text processing. GA is deployed to text processing as an optimization problem. GA is used in text clustering, Text classification and Text Automatic summarization [14].

Goal of an automatic text summarization system is to generate a summary of the original text that allows the users to obtain the main pieces of information available in that text, but with a much shorter reading time [3]. In addition, an important data preprocessing task for effective classification is the attribute selection task, which consists of selecting the most relevant attributes for classification purposes [9]. So that term weighting method, feature selection and ranking are important task in text mining process that can be done upon TF*IDF method; Information Gain Ratio; Mutual Information. The goal of this paper is to deploy Genetic Algorithm in term weighting and feature selection in text mining process. GA has been used in text processing. The rest of the paper is organized as follow: in the next section we will review the related work on term weighting, information extraction, concept distribution and Genetic Algorithm. In the third section we state our algorithm for using Genetic Algorithm in topic identification and in the fourth section we will evaluate our method and some remarks about our method will be explored in last section.

S. M. Khalessizadeh, Department of IE, Sharif University of Technology, Tehran, Iran (e-mail :khalessi@sharif.edu)
R. Zaefarian, Department of IE, Sharif University of Technology, Tehran, Iran (corresponding author: rzaefarian@ie.sharif.edu)
S.H. Nasseri, Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran (e-mail: nasseri@math.sharif.edu).
E. Ardil, Department of Computer Engineering, Trakya University, Edirne Turkey (e-mail: ebruardil@trakya.edu.tr).

II. Related Work

A. Topic identification and concept extraction

Brown and Yue (1983) pointed out that there are two kinds of topics: one is sentence topic and the other is discourse topic. The discourse topic is usually the form of topic sentence [4]. In our approach, we use the term "discourse topic" as a representation of the document content in order to distinguish it from other documents in corpus. Discourse topic should be human understandable topic to allow a person to quickly see what the topic is about [6]. The foundation of the topic identification process is frequent item sets [6]. In our case, a frequent item set is a group of individual or combination terms that occurs together in a document. We call combination terms in topic "concept" and use it to distinguish from an individual term, simple keyword or term with the surface expression [12]. In our approach, "concept" is used to state some related words that point to a specific entity or impression in a document. By using controlled vocabularies, the concept set of documents could be extracted from a text [2]. Also automated learning of document sets can be used. Extracting concept sets by using a dictionary will lead to domain dependency in determined concept sets.

We denote topic of a document with Td_i and we can represent topic of document I as a combination of concepts or terms that are identified by text mining process. The weight of concept/ term j in document i will be denoted as Wd_ic_j and we can represent document i as a combination of concept/term weight like Wd_i:

$$Wd_i : \{ Wd_ic_1; Wd_ic_2; Wd_ic_3 ; \ldots \},$$

where Wd_ic_j means the weights of concept/term j in document i. Few researchers would claim that a word representation is optimal, but the difficulty of automated natural-language understanding has limited our ability to use a richer representation scheme [3]. Because of same keywords may have some different meaning, and in inverse case, some different keywords, refer to same meaning, using of concept in topic identification is more useful than term.

Open Science Index, Mathematical and Computational Sciences Vol:2, No:1, 2008 publications.waset.org/10232/pdf

B. Concept/Term Distribution

The distribution of concepts or term in a document is an important factor to be taken into account. The discrimination of the distribution of the two concepts in two documents will diminish the similarity between two documents. Weng and et.al [15] proposed a common statistical equation, 'standard deviation', to represent the dispersion of the concepts in a document. They divided standard deviation by maxline(d_i) to normalize standard deviation by the size of each document.

$$\text{Vard}_{ic_j} = \sqrt{\frac{\sqrt{x^2} - (\sum x)^2 / n}{(n-1)\max line(di)}} \qquad (1)$$

Here x indicates the position where the phrase of a concept appears in a document, n denotes the total phrase number in a document, maxline(d_i) denotes the maximum line of document i and width of a line is set to a constant in order to fix the width of every article [15]. Vard_{ic_j} is the standard deviation of the concept position in the document. We use standard deviation to represent the distribution of the concept in a document. The higher the standard deviation of the document is, the greater the dispersion in the distribution.

We use Weighted Topic Standard Deviation to state the concentrate of document on its identified topic. Assume that Wdi :{ Wdic1; Wdic2; Wdic3 ;…}, express the weight of each concept or term in topic of document. The WTV of document i will be calculated with following formula:

$$\text{WTSD(di)} = \sqrt{\sum_{j,k} \frac{wd_i c_j (x_{j,k} - \bar{x}.wd_i c_j)^2}{(n-1).\overline{w}.\max line(d_i)}} \qquad (2)$$

where $wd_i c_j$ represent weight of concept or term j in document i and $x_{j,k}$ represent the number of line that concept j accrued. $\overline{w}$ represents the average of weight of concepts or terms in document i. As discussed earlier, the small topic standard deviation, states the more concentrate on topic.

C. Genetic Algorithm

Genetic algorithms are heuristic optimization methods whose mechanisms are analogous to biological evolution [16]. A good general introduction to genetic algorithms is given in [17]. In Genetic Algorithm, the solutions are called individuals or chromosomes. After the initial population is generated randomly, selection and variation function are executed in a loop until some termination criterion is reached. Each run of the loop is called a generation. The selection operator is intended to improve the average quality of the population by giving individuals of higher quality a higher probability to be copied into the next generation. The quality of an individual is measured by a fitness function.

III. Genetic Algorithm Application in Topic Identification

A. Problem Description

As mentioned before, there are many concepts and terms in document that can be recognition by part of speech technique and concept and information extraction. Some of them contribute in the main topic of document. Our main problem is to distinguish the weight of each concept or term in topic of document. We represented the weight of concepts and terms as Wd_i vector in previous section. Each concept or term can have a weight between 0 and 1. For simplifying our problem, we can consider weights as a binary number. That means the related concept or term belongs or doesn't belong to topic of document. A chromosome is defined as a list of concept or term weights which have real or binary numbers. The definition of a chromosome is represented as $J = (j_1, j_2, \ldots, j_i, \ldots, j_L)$, where j_i denotes the weight of the concept i and L is the number of concept to be considered. Each gene represents a concept or term weight. The genes of initial chromosomes are generated randomly and the range of weight values is from 0.0 to 1.0 for experiments.

B. The Fitness Function

We use WTSD as our fitness function. WTSD measures the performance of weighting precision. As we discussed in previous section, the small WTSD state much more concentration on topic at document. For each solution (individual or chromosome in the generation) we can calculate the WTSD with using eq. 2. In each solution the weight of each concept or term is determined and so that we can consider each solution as a vector of topic of document. Vectors with small WTSD show the concentration of document on that topic. Based on this fitness function, we can select the appropriate solutions as a parent for offspring. We use truncation selection, where the parents are selected randomly from a half of the population in the decreasing order of quality [15].

C. Genetic Operators

The genetic algorithm uses crossover and mutation operators to generate the offspring of the existing population. Before genetic operators are applied, parents have been selected for evolution to the next generation. We use the crossover and mutation algorithm and produce next generation. The probability of deploying crossover and mutation operators can be changed by user. In all of next generation, WTSD has used as our fitness function.

D. End Condition

GA needs an End Condition to end the generation process. If we have no sufficient improvement in two or more consecutive generations; we can stop the GA process. In other

Open Science Index, Mathematical and Computational Sciences Vol:2, No:1, 2008 publications.waset.org/10232/pdf

cases, we can use time limitation as a criterion for ending the process.

E. Our Algorithm

According to the above sections, our algorithm is explained in the following:

1. [Start] Generate random population of n chromosomes (suitable solutions for the problem that is explained in section A)
2. [Fitness] Evaluate the fitness f(x) of each chromosome x in the population with WSTD fitness function (section B).
3. [New population] Create a new population by repeating following steps until the new population is complete
    1. [Selection] Select two parent chromosomes from a population according to their fitness.
    2. [Crossover] With a crossover probability cross over the parents to form new offspring.
    3. [Mutation] With a mutation probability mutate new offspring at each locus.
4. [Accepting] Place new offspring in the new population for a further run of the algorithm.
5. [Replace] Use new generated population for a further run of the algorithm
6. [Test] If the end condition (section D) is satisfied, stop, and return the best solution in current population, otherwise go to step 2.

IV. Experiments and Results

As a test for our genetic algorithm, we applied it for concept weighting in a standard text in Persian language. Existence of various numbers of historical, religious, scientific and literary texts in Persian language has increased the necessary of effective tools for processing texts. Many of these texts are our national treasure and their detail analysis is essential. The increasing volume of texts in Persian language reveals the importance of employing a modern technique for their process.

In many languages there are some software packages for text mining and text processing. However, because of structural differences between Persian and other languages, current software packages in the world are not useful in processing Persian texts. After developing algorithms and software packages which can process Persian texts we use our Genetic Algorithm in our package and try to identify the topic of documents based on concept distribution. We deploy our algorithm in a set of literary texts. First of all, for each document, a list of concept was produced by using one of the following two approaches: (1) an automatically-generated concept list, this kind of concept list is called an "automatic list". (2) A manually-generated concept list, produced by a Persian teacher by selecting the most relevant concept of the text. This is called a "manual list".

We use both TF*IDF and GA methods to identify the concept list for each document. We use following measures as a standard measure to address the performance of our algorithm:

$$PRECISION = \frac{Concept - found - and - correct}{total - concept - found}$$

$$RECALL = \frac{Concept - found - and - correct}{total - concept - correct}$$

The result is addressed in the following table:

TABLE I

Number of word in document	Number of concept in manual list	Automatic Generated List			
		Precision in GA	Recall in GA	Precision in TFIDF	Recall in TFIDF
1-100	10	46%	60%	43%	46%
101-150	12	53%	67%	44%	47%
150-200	11	50%	64%	57%	57%
>200	13	50%	69%	50%	56%

As shown in above table, the GA has a notable improvement compared to traditional TF-IDF. The accuracy of both methods is lower when the size of the documents is small.

V. Conclusion

As mentioned earlier, Concept weighting and topic identification is an essential task for document management. Most of the past research focused on TFIDF algorithms. We propose a new algorithm, based on concept distribution and by using Genetic Algorithm, to identify the weight of concept. According to the empirical evaluation result, the proposed technique was more effective than the traditional TFIDF method. It should be considered that this algorithm can be improved. Our algorithm's largest precision and recall rate, reported in above tables, were 53% and 69% respectively. So that we can expect that the performance of algorithm can be improved by changing in some parts of algorithm.

REFERENCES

[1] T. Anand and G. Kahn, "Opportunity explorer: Navigating large databases using knowledge discovery templates", In Proceedings of the 1993 workshop on Knowledge Discovery in Databases.
[2] C. Apte, F. Damerau and S.M. Weiss, "Automated learning of decision rules for text categorization", ACM Transactions on Information Systems, 12 (1994) 233–251.
[3] C. Blake, W. Pratt, B. Rules and F. Features, "A Semantic Approach to Selecting Features from Text'', ICDM, (2001) 59–66.
[4] G. Brown and G. Yule, Discourse Analysis. Cambridge University Press, 1983.
[5] S. Chakrabarti, ''Data mining for hypertext: a tutorial survey", ACM SIGKDD explorations, 1 (2000) 1–11.
[6] C. Clifton, R. Cooley and J. Rennie, T. Cat, ''Data mining for topic identi_cation in a text corpus'', 3rd European Conference of Practice of Knowledge Discovery in Databases, Prague, Czech Republic, 1999.
[7] K. Ezawa and S. Norton, "Knowledge discovery in telecommunication services data using Bayesian Models", In Proceedings of the First International Conference on Knowledge Discovery (KDD-95), 1993.
[8] W. Fan, M.D. Gordon and P. Pathak, "A generic ranking function discovery framework by genetic programming for information retrieval", Information Processing and Management 40 (2004) 587–602.
[9] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers, 1998.
[10] I. Mani and M.T. Maybury, Advances in Automatic Text Summarization, MIT Press, 1999.
[11] T. W. Manikas and M.H. Mickle, ''A genetic algorithm for mixed macro and standard cell placement", 27th ACM IEEE Design Automation Conference.
[12] T. Nasukawa and T. Nagano, ''Text analysis and knowledge mining system", IBM SYSTEMS JOURNAL, VOL 40, NO 4, 2001.
[13] S.N. Sancheza, E. Triantaphylloua, J. Chenb and T. W. Liaoa, "An incremental learning algorithm for constructing Boolean functions from positive and negative examples", Computers & Operations Research 29 (2002) 1677–1700.
[14] C.N. Silla, G.L. Pappa, A. Freitas and C.A. Kaestner, "Automatic text summarization with genetic algorithm-based attribute selection", 9th Ibero-American Conference on AL, Lecture Notes in Computer Science, 3315 (2004) 305–314.
[15] S.S. Weng, Y.J. Lin and F. Jen, ''A study on searching for similar documents based on multiple concepts and distribution of concepts", Expert Systems with Applications 25 (2003) 355–368.
[16] M. Mitchell, An Introduction to Genetic Algorithm, MIT Press, 1996.
[17] G.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley, New York, 1989.

Open Science Index, Mathematical and Computational Sciences Vol:2, No:1, 2008 publications.waset.org/10232/pdf