Journal of Rajasthan Academy of Physical Sciences
ISSN : 0972-6306; URL : http://raops . org. in
Vol.13, No. 2, June, 2014, pp. 171-179

AUTOMATIC TEXT SUMMARIZATION USING PAGE RANK AND GENETIC ALGORITHM*

SHASHANK GUPTA, ANUSHREE JAGRAWAL and NEHA MATHUR

Birla Institute of Technology, Mesra, Jaipur Campus
27, Malviya Industrial Area, Jaipur 302017, India.

Email: 27392shashankgupta@gmail.com,
anshi31jagrawal@gmail.com, nhmathur9@gmail.com

Received: April 3, 2014

Abstract: To extract "only" important 'Information' from data a summarizationsystem is used. An automatic text summarization system using extractive method is architected which learns from data itself and extracts important information according to user's requirement. Unsupervised form of machine learning has been employed, inspired by Google's page rank algorithm which constructs a network of sentences and then ranks the pages according to their relative importance. The document is formulated as Markov chain[1]and then ranking is done based on probability of transition to other sentence using random walk approach in case of small document. In case of large documents, the solution space becomes very large so a simple optimization algorithm has chance of trapping in local optimal point, so to deal with it genetic algorithmhas been applied which guarantees global optimum solution.

Keywords: Markov chains, page rank, abstractive, extractive, genetic algorithm.

2010 Mathematics Subject Classification: 60Jxx, 65Cxx, 90B15, 05C81

---

*This paper was presented in COMPUTATIA III held at V I T, Jaipur, India during 28th and 29th Nov., 2013.

1. Introduction

Automated Text summarization system is a system which takes a structured data and gives output as summary of the text which contains main information of the text. It derives its working principle from the way the human's summarize a document or text. Humans tend to include that information from document which is relevant or more important by checking the importance of sentences and including most important sentences. In the similar way automatic text summarization system works. It gives each sentence some importance factor based on different features which are selected based on the type of summary to be expected. This task of summarization comes under the category of information retrieval, which itself is a form of data mining and machine learning. Here information is extracted from data through different machine learning approaches. Broadly automatic text summarization is classified into two categories: Extractive summarization and abstractive summarization[7]. Extractive summarization works by simply weighting sentence based on different features and then including those sentences which have maximum weights. Features are selected on the basic of different types of summaries like query based summary, content based summary etc. In abstractive summary individual words of different sentences are weighted and the most important words are selected and new sentences are constructed using natural language processing techniques. It requires a sound knowledge in languages grammar and very efficient natural language techniques like hidden markov model, parts of speech tagger etc. So abstractive summarization is not so developed area and not much systems have been developed on it. So in this paper we use extractive summarization. A general algorithm for such systems is to weigh sentences as weights $w_1,w_2,w_3.....w_n$ and then using linear combination of such weights to find central weight as

$$W = w_1 + w_2 + w_3 + \ldots + w_n$$

2.sAlgorithm

The algorithm used is: first for small data set i.e. for short documents, it is a stochastic graph based method for computing relative importance of textual units. In traditional methods sentences are weighted relative to a central pseudo sentence [6], but this does not lead to a very efficient algorithm, so in this system weighting is done using Eigen vector based centrality in graph representation of sentences. This ranking is called Lex rank (Lexical page rank) derived from extremely popular Google's page rank algorithm [3] which is a problem to model behavior of a random surfer. This method is more efficient compared to more popular centrality based models and also is more susceptible to noise in data. Ranking is done based on a feature called IDF (inverse document frequency)[2]. It is very useful parameter in information retrieval systems like keyword extraction systems. If we use normal frequency of words in a document we will get somewhat normal distribution of words. So to calculate relative importance we use a new feature called collection frequency. Some terms appear too often in documents like word auto would appear too often in a document of an auto industry. So concept of collection frequency is used for attenuating the effect of terms that occur too often as they are not meaningful for relevance. So idea is to scale down the terms weights of terms with high collection frequency. IDF is defined as

$$IDF(I) = \log\left(\frac{N}{TF(I)}\right)$$

where N is number of sentences and TF is number of sentences in which word I occur. Now it is also multiplied by the frequency of the word I. Now this feature is used to weigh sentences. After weighing sentences they are represented as vectors and the document is represented as a vector space. Now to calculate similarity between sentences concept of cosine similarity is used. Cosine similarity between two vectors **A** and **B** is defined as

$$\cos(A,B) - \sin(A,B) = \frac{A.B}{(|A|.|B|)}$$

Similarly for vectors in documents vector space cosine similarity is defined and a similarity matrix is constructed based on following formula.

$$idf - modified - cosine = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}}$$

Now this similarity matrix is basically adjacency matrix for a graph with node as sentences and edges as similarity between them. Or edges can be represented as sentences coming in order next to the other. So this graph can be modeled as a Markov chain, and random walk in this graph will result in a sequence of sentences to be included in summary. Now after modeling graph as a Markov chain the Lex rank of vectors is calculated which is calculated by random walk in graph. Now if cycles exists then this random walk can result in an infinite loop so some damping factor is attached which prevents the loop from entering infinite state. This formula for lex rank is derived from Google's page rank [3] algorithm.

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{p(v)}{\deg(v)}$$

where p(u) is the probability of random user visiting a web page u, and d is the damping factor. In this context it is the relative importance of a sentence, Cosine similarity is used in this calculation. Following algorithm summarizes the Lex Rank calculation.

1 **Input:** An array S of n sentences, cosine threshold t
2 **Output**: An array L of Lex Rank scores
3 Array Cosine Matrix[n][n];
4 Array Degree[n];
5Array L[n];
6 for i ← 1 to n do

```
7            for j ← 1 to n do
8            CosineMatrix[i][j] = idf-modified-cosine(S[i],S[j]);
9                        if CosineMatrix[i][j]> t then
10                       CosineMatrix[i][j] = 1;
11                       Degree[i]++;
12                       end
13                       else
14                       CosineMatrix[i][j] = 0;
15                       end
16           end
17end
18 for i ← 1 to n do
19            for j ← 1 to n do
20           CosineMatrix[i][j] = CosineMatrix[i][j] / Degree[i];
21           end
22 end
23 L = PowerMethod(CosineMatrix,n,ϱ);
24 return L;
```

Algorithm: Computing LexRank scores.

After the Lex Rank is calculated, now sentences can be sorted according to their Lex rank and displayed to user according to his requirements. Now Genetic algorithm [5] has been employed to find optimal solution of this network graph which optimizes the objective function as linear combination of some parameters as specified below. Genetic algorithm's parameters are usual with selection as roulette wheel, then crossover with a hybrid approach used: one half of chromosome randomly shuffled and other half swapped with other parent chromosome to maintain the parent's traits. Then mutation operation is used with some mutation probability. The cycles of population is chosen till it converges to an optimal value

3. Parameters used for fitness

To calculate fitness of candidate from candidate population fitness function has to be defined. Fitness function is defined as weighted sum of different parameters chosen. The parameters are:

1. Lex rank: it is the Eigen vector of the similarity as vectors and then cosine similarity matrix of the sentences. Sentences are defined between them is defined. This similarity matrix is used to define a graph representation of the document. Then using random walk concept ranking is done.
2. Position of sentence: In a statistical study it was observed that sentences with position as starting of a document meaning sentences in first and till second paragraphs are considered to be more relevant to users, i.e. they tend to pay more attention to starting sentences and also sometimes last sentences, so position of sentence is also considered a candidate for selection is fitness function.
3. Numerical data: In a document sentences having numerical data are considered more relevant as compared to sentences which don't have numerical data. For example in an automobile's documents sentences having numerical information like displacement of engine(it's volume), its torque like 100NM, power like 1000 BHP. These are more relevant so given more importance.
4. Similarity with title: Sentences having some similarity with title will be considered to be more important for obvious reasons. For instance in a document of computer networks sentences having words network and computer will be more relevant as they express theme of the document.

4. Results

To verify the results following document is considered:

1. Iraqi Vice President TahaYassin Ramadan announced today, Sunday,

2. that Iraq refuses to back down from its decision to stop cooperating

3. with disarmament inspectors before its demands are met.

4.Iraqi Vice president TahaYassin Ramadan announced today, Thursday,

5.that Iraq rejects cooperating with the United Nations except on the

6.issue of lifting the blockade imposed upon it since the year 1990.

7.Ramadan told reporters in Baghdad that "Iraq cannot deal positively

8.with whoever represents the Security Council unless there was a clear

9. stance on the issue of lifting the blockade off of it.

10. Baghdad had decided late last October to completely cease cooperating

11. with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work

12. became very limited since the fifth of August

Here each sentence is number as Si, 1<=i<=11.

The Lex score of each sentence with threshold 0.1 is.

Sentence Number	LR(0.1)
S1	0.6
S2	0.84
S3	0.34
S4	0.75
S5	0.59
S6	0.74
S7	0.31
S8	0.89
S9	0.51
S10	0.61
S11	0.55

Now this vector can be sorted as specified in the introduction this vector is combined by other features (Similarity with title etc.).

Now this linear combination is then searched for optimum solutions and then sentences are displayed according to their score sorted from maximum to minimum.

5. Conclusion and scope

Automatic text summarization system is a complex system mainly due to different views of summary for different persons. But its scope is very vast. Talking about its scope, it can be employed in many areas with different purpose. For example in computer network information flows in form of packets, but even with huge advancements in signal flow techniques signals today can be transferred with gigabits of speed, but still the actual bandwidth an end user gets is still less, main reason being packets stopped and analyzed at every hop(each router in its path), routing information send by routers to each other which consumes bandwidth of cpu of router. Basically a packet is forwarded by a router through its routing table which contains next hop information; this routing table is updated by neighbors which send information about every path. So this is tedious and time consuming path as routers are busy updating their routing table. So to avoid this time delay at routers, route summarization technique is used, it basically summarizes routers topology database(database of routes), so an efficient summarization technique will speedily and efficiently summarizes routes and will save router's cpu bandwidth and that bandwidth can be utilized in forwarding packets so overall bandwidth of network can be improved.

Another interesting application is in context of semantic web [4]. Current web structure is only hyperlinks of web pages with databases connected. The concept of semantic web is an intelligent web through which we can extract "information" based on query system. It can be basically build by analyzing data from web pages and developing a rule base or knowledge base from it through first order logic and data mining techniques. So It would require to analyze tons

of data, so to make this process more efficient data from web pages can be summarized so that only relevant information remains then we can apply NLP and information retrieval techniques to it. This is only way to make semantic web. Wolfram alpha is an excellent example of semantic web at a prototype stage.

Also this technique can be employed to any area with large data but whole data need not be considered only relevant data is of computational importance.

6. References

[1] Barzilay, R. and Elhadad, M. (1999). *Using Lexical Chains for Text Summarization, MIT Press Masschusetts, London, England.*

[2] Knuck G, (2004) *Tim Berners-Lee's Semantic* Web ISSN 1560-683X, Published by InterWord Communications for the Centre for Research in Web-based Applications, Rand Afrikaans University

[3] Larry, P., Sergey, B., Motwani, R. (1998) *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab.

[4] Manning, C.D., Raghavan, P. and Schütze, H. *(2008) Information Retrieval*, Cambridge University Press.

[5] Mitchell, M.(1999) *An introduction to Genetic Algorithm,"*A Bradford Book", MIT Press, Cambridge, Massachusetts. London, England

[6] Radev, D. R., Jing, H., and Budzikowska, M. (2000).*Centroid-based summarization of multiple documents: sentence extraction, utility based evaluation, and user studie. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization,NAACL-ANLP-Autosum 2000.*

[7] Witbrock, M.J. and Mittal, V.O. (1999) *Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries, In* Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)