



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirce.com

Vol. 4, Issue 12, December 2016

Text Summarization: Ranking Sentence from Key Phrase Extraction

Vidya Panchal¹, Prof. A. Anjum²

M.Tech Student, Department of Computer Science and Technology, Maharashtra Institute of Technology,
Aurangabad, India¹

Assistant Professor, Department of Computer Science and Technology, Maharashtra Institute of Technology,
Aurangabad, India²

ABSTRACT:Text Summarization involves reducing large amount of text into smaller in size. So that we can get information about what kind of the data are present in that document. We can get more correct amount of information of the large text or document. Now days, various techniques are available to obtain the summary from the document or large text. In this, we represent the technique that help you to get summary of the document. This technique based on the ranking the sentence from key phrase extraction algorithm.

KEYWORDS: Ranking, Key phrase, text summary, Data mining algorithm.

I. INTRODUCTION

Now a day's tremendous amount of the data available everywhere in the form of wiki website, large text, documents and internet. So that to read that much amount of the data is the time consuming. Text summarization helps us to summarize that data into in the form of abstraction so that we can get exact and generic and more specific amount of information of that document or internet website or large text. According to MichelangeloCecia ,CorradoLogliscia , LucreziaMacchia[1] , they stated thatsummary of the document image is obtain using keyphrase extraction algorithm. According to Jackson, P., they stated that summary can be obtained in the form of extracts, they are formed by extracting the key phrases of the original text or document and abstracts of the document. When they are formed by concluding the meaning of the original document or by re-generating the content of original document.[2] According toSparck Jones K., They stated that summary can be obtained by more extractive way like selectingmore salient sentences through Machine Learning or Data Mining algorithms. So that we can obtained the summary of the original document by more effective way. According to PaiceC.D[3]. , they stated that sentences are mentioned in the form of lexical and structural features (e.g., keywords frequency, title keywords, sentence location, indicator phrases, etc.) and described or represented by as vectors of quantitative and categorical measures of those features (attribute-value representation) in the original document. And Zhuli, X., Li X., Barbara, D.E., Peter, N., Weimin, X., Thomas T[4] , they mentioned that the summary of the document is also obtained by Several supervised learning techniques which are the part of the Data Mining algorithms. To rank new sentences of the original document final ranking sentence function is used. Svore et al.[9] stated that resort to a neural network pair to generate a ranking composed of three sentences is also used to developed summary of the original document.

II. PROPOSED ALGORITHM

A. SENTENCE RANKING FOR SUMMARY:

Learning a preference model that leads to identify preference relations to be applied to new documents is used to solved the problem of calculating summary of the Document Text or URLs.

i. Mining Preference relations:

The problem of mining preference relations between sentences can be obtained as follows.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 4, Issue 12, December 2016

Given:

- A database schema S with h relational tables $S = \{T_1, T_2, \dots, T_h\}$
- Two sets PK and FK of primary and foreign key constraints on tables in S
- A target relation $T \in S$ representing sentences that play the role of reference objects
- A precedence relation $PT \in S$ with two attributes. Each tuple in this table represents an ordered pair of reference objects where the first reference object precedes the second one.

Find: A probability estimation $P(a < b | a, b)$ for any couple of sentences a and b which are present to a new document represented according to the schema $S - \{T, PT\}$. Objects in $S - \{T, PT\}$ play the role of task relevant objects, while the

Precedence relation implicitly defines a partial ordering between two sentences.

In our approach, it is also possible to avoid to consider some sentences belonging to parts of the document that are not considered relevant for the task at hand (e.g. sentences in tables or sentences in references of a original paper). This means that the preference relation in training data does not necessarily express total ordering of sentences in training documents.

By applying the Bayes theorem, $P(a < b | a, b)$ can be computed as:

$$P(a < b | a, b) = P(a < b)P(a, b | a < b) / P(a, b) \dots \dots \dots (1)$$

where:

- $P(a < b)$ in (1) denotes the prior probability that a sentence precedes another. This probability might be different from 0.5.
- $P(a, b) = P(a < b)P(a, b | a < b) + P(b < a)P(a, b | b < a)$.

ii. Patterns construction:

The relational pattern discovery is obtained by exploring level-by-level the lattice of relational patterns ordered according to a generality relation ($\succ=$) between patterns. Given two patterns P_1 and P_2 , $P_1(\succ=)P_2$ shows that P_1 (P_2) is more general (specific) than P_2 (P_1). Hence, the search gives us the most general pattern for the summary and iteratively alternates the candidate generation and candidate evaluation phases to obtain the summary. In [14], the authors mention an enhanced version of the level-wise method [15] to discover patterns from data in multiple tables of a relational database for the calculating the summary of the document. The space of linked relational patterns are searched by candidate patterns, which is constructed according to the θ -subsumption generality order [16].

This shows possible to obtain a levelwise exploration of the lattice of relational patterns ordered by θ -subsumption. Specifically, patterns are obtained by forming the pattern space one level at a time starting from the most specific Pattern (the pattern have only the *preference/2* predicate) and then by applying a breadth-first evaluation in the relational patterns ordered according to $\succ=\theta$.

iii. Ranking Reconstruction

The main goal of this step is to produce ranking the sentences the sentences for the construction of the summary of the original document.

For the ranking the sentence for construction of the summary following algorithm are used:

Algorithm 1 ranking identification algorithm:

- 1: find ranking ($G = [V, E]$): Ranking L
- 2: $L \leftarrow \emptyset$;
- 3: while ($\#L < \#V$) do
- 4: $L.add(\arg \max_{(b_i \in V/L)} SUMPREFG(b_i))$;
- 5: end while



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirce.com

Vol. 4, Issue 12, December 2016

Where,

- $G = \{V, E\}$ be a *labeled* directed graph where $V = \{b \in T\}$ and $E = \{(a, b, w(a,b)) \in V^*V \times [0, 1] | w(a,b) = P(a < b | a, b)\}$ is the set of weighted edges of the graph where weights are the probabilities $P(a < b | a, b)$ computed according to equation (1),
- $SUMPREFG: V \rightarrow [0, \#V]$.

This Algorithm works on the method for the ranking identification. If the $SUMPREFG()$ of the sentence is high then that sentence is added to the summary of the document text. Once this implementation of the algorithm is done then that 'n' sentences is obtained from the algorithm is added to the Summary of the original document.

III. DATA EXTRACTION AND REPRESENTATION

Extraction of the key phrases from the document involves tokenization, sentence splitting, part-of-speech (POS) tagging, stop-word removing and stemming.

The representation of the sentences form by the ranking the key phrases extracted by extraction algorithm is done by means of a phase of natural language processing.

IV. EXPERIMENTS

Dataset	WISDOM++	GREEDYEXP	GREEDYUNIFORM	SVD	FURTHEST
TPAMI	80.20	77.43	77.52	78.61	80.10
ICML	59.05	57.23	58.35	51.43	53.45

1. Cosine similarity observed between the abstracts and the system's summaries of the document

Dataset	WISDOM++	GREEDYEXP	GREEDYUNIFORM	SVD	FURTHEST
TPAMI	0.90	0.89	0.88	0.90	0.58
ICML	0.80	0.85	0.85	0.78	0.41

2. Rouge-1 observed between the abstracts and the system's summaries of the document

The result of the first table shows the performance of the system and the result of the second table shows behaviour comparable with the best competitors.

V. CONCLUSIONS

In this paper, we purpose the how to obtained the summary from large amount data or large document or URLs. For this we use sentence ranking algorithm to form summary of the original document using the key phrases. Key Phrases for ranking sentences is obtained by key phrases extraction algorithm using relational data mining algorithm or tools.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 4, Issue 12, December 2016

REFERENCES

1. Michelangelo Cecilia,*, Corrado Logliscia, Lucrezia Macchia. "Ranking Sentences for Keyphrase Extraction: A Relational Data Mining Approach" *Procedia Computer Science* 38 (2014) 52 – 59
2. Jackson, P.. The oxford handbook of computational linguistics edited by ruslan mitkov. *Computational Linguistics* 2004;30(1):103–106.
3. Sparck Jones, K.. Automatic summarising: The state of the art. *Inf Process Manage* 2007;43(6):1449–1481.
doi:<http://dx.doi.org/10.1016/j.ipm.2007.03.009>
4. Paice, C.D.. Constructing literature abstracts by computer: Techniques and prospects. *Inf Process Manage* 1990;26(1):171–186.
5. Dzeroski, S., Lavrač, N.. *Relational Data Mining*. Springer-Verlag; 2001
6. Ceci, M., Loglisci, C., Ferilli, S., Malerba, D.. Project d.a.m.a.: Document acquisition, management and archiving. In: Agosti, M., Esposito, F., Meghini, C., Orio, N., editors. *IRCDL*; vol. 249 of *Communications in Computer and Information Science*. Springer. ISBN 978-3-642-27301-8; 2011, p. 115–118.
7. Ceci, M., Berardi, M., Malerba, D.. Relational data mining and ILP for document image understanding. *Applied Artificial Intelligence* 2007;21(4&5):317–342.
8. Ceci, M., Berardi, M., Porcelli, G., Malerba, D.. A data mining approach to reading order detection. In: *ICDAR*. IEEE Computer Society. ISBN 978-0-7695-2822-9; 2007, p. 924–928.
9. Svore, K., Vanderwende, L., Burges, C.. Enhancing single-document summarization by combining ranknet and third-party sources. In: *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics; 2007, .
10. Helft, N.. *Progress in Machine Learning*; chap. Inductive generalization: a logical framework. Sigma Press; 1987, p. 149–157.
11. Ceci, M., Appice, A., Malerba, D.. Emerging pattern based classification in relational data mining. In: Bhowmick, S.S., K'ung, J., Wagner, R., editors. *DEXA*; vol. 5181 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-540-85653-5; 2008, p. 283–296.
12. Robinson, J.A.. A machine oriented logic based on the resolution principle. *Journal of the ACM* 1965;12:23–41
13. Ceci, M., Appice, A., Malerba, D.. Discovering emerging patterns in spatial databases: A multi-relational approach. In: Kok, J.N., Koronacki, J., de M'antaras, R.L., Matwin, S., Mladenic, D., Skowron, A., editors. *PKDD*; vol. 4702 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-540-74975-2; 2007, p. 390–397.
14. Mannila, H., Toivonen, H.. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery* 1997;1(3):241–258.
15. Kamishima, T., Akaho, S.. Learning from order examples. In: *ICDM*. IEEE Computer Society. ISBN 0-7695-1754-4; 2002, p. 645–648.
16. Liu, K., Terzi, E., Grandison, T.. Manyaspects: a system for highlighting diverse concepts in documents. *PVLDB* 2008;1(2):1444–1447.