Intricacies of an Automatic Text Summarizer

Lincy Meera Mathews [#1], Dr E Sathiyamoorthy [*2]
[#]Department of Information Science & Engineering, MSRIT
Bangalore
[1] lincymm99@gmail.com
[*] Associate Professor, School of Information Science & Technology
Vellore University, Vellore
[2] esathiyamoorthy@vit.ac.in

***Abstract*— The creation of an abstract over a text document prepared by a computer program is defined as an Automatic Text Summarizer. This abstract of the text document must however contain all the salient features of the original document. This paper tries to cover the necessary functional modules that complete an automatic text summarizer. It also highlights the trends and challenges in text summarization. Surveys of certain text summarization techniques are also mentioned.**

Keywords: Text summarizer, Natural language processing, Extract, Abstract, Hybrid

I. Introduction

Human Beings have now access to abundance of information on the net. However access to the relevant information required by the user is still a challenge today. In view of this, text summarization techniques[1] is now one of the most important and well researched tool for assisting and retrieving of digital information.

Text summarizer was mainly developed to

1) Improve the quality of text classification techniques such as classification, clustering and regression. The output of classifiers is highly dependent on the quality of summarized text document.
2) Reduce time spent by researchers, academics by access to quality abstracts of digital documents.
3) Access to relevant and important facts immediately. Humans have the tendency to oversee important and critical facts or sentences. However a text summarizer will automatically cover the important facts of the document.

Section 2 presents an outline and background in the area of text summarization. Section 3 investigates the methodologies and relevant modules required in a text summarizer. Section 4presents a survey of text summarization tools with their techniques. Section 5 covers the challenges of text summarizer. Section 6 concludes the paper.

II. ABOUT TEXT SUMMARIZER

A. Definition and Aim:

Hovey,E.H [2] defines summary as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s) and that is no longer than half of the original text(s).

The goal of a text summarizer program is to summarize a text document by

1) Distilling the most important facts in the respective document.
2) Covering all the salient aspects and topics of the document.
3) Narrowing down on the most precise and complete statement to represent a topic, paragraph or a sentence.
4) Inclusion of only the most necessitated statements without redundancy, ambiguities and error in representation of information.

The structure of a Text document will necessarily contain the following: Title of the document and sentences that are formed by concatenating a group of words. However the document may or may not contain the following: Subtopics also called as sub headers, paragraphs defined as connected sequences of sentences and conclusion. Tapping on each unit of the document, the text summarizer must succeed on summarizing by using the information retrieved and formulate the best represented sentence for the document.

Summarized document need not be a direct function of the text document alone. It can be a bi function of text document and user knowledge. The summarization must be dependent on the knowledge depth of the user. If the user is a novel, he gets a simplified version of the document as he might not be aware of the jargons used. A knowledgeable user might be interested in the core area of the topic being covered in the document.

A summarized document is characterized by its

1. The condensation rate or the reduction rate. The length of the text document divided by the abstract or summary length is defined as the condensation rate. The rate is to be less than 1.
2. The quality of information produced in the summary is another factor that characterizes the summary. The degree of relevant information extracted from the text document with respect to the user's knowledge indicates the quality of summary.
3. The degree of well formalness: Grammatically and structurally the summary must be well formed

B. Types of Summarizer:

Stastical approach as wells as linguistic approaches are the two ways of carrying out Text summarization. Linguistic approach works on domain knowledge whereas the stastical approach relies on machine learning techniques. The product of a summary system can be an abstract, extract, capsule or a hybrid.

1. Extract: One of the simplest and easy to implement technique. The main work involves identification of key statements and paragraphs. These sentences are then used to form the summarization.
   The significance of a sentence can be evaluated by
   a. Following a bottom –up approach for information retrieval.
   b. Extraction of sentence if presence of cue words in the respective sentence
   c. Analysis of individual and surface level word and frequency which are identified as relevant features

The extraction process is a two-step phase.

Phase 1 identifies the end of sentences by presence of dots. The stop words are then eliminated from the document while care is taken to remove only those stop words that do not involve any ambiguities and their removal does not cause change or loss in information. Stemming is then performed to uproot the stem of each word.

Phase 2 identifies the salient features of a document. Various Feature selection methods are used to choose the representing features of the document. Weights are then assigned using the TD-IDF feature weight or using the Mutual information, Chi-squared criteria. The respective weight of sentence is calculated as sum of weights of thematic features. Highly weighted sentence are included in the abstract. Presence of headline words also adds to the weight age of the sentence.

The extractive method usually employs machine learning techniques like neural networks, naïve baye's classifier. Neural Summarizer (NeuralSumm) [4] is an automatic text summarizer that is based upon neural network. During the training phase, a model classifier is created where it becomes capable of identifying relevant sentences for inclusion in extract. Classy [5] summarization project employs a classification system using Machine Learning approach, in which the classification task is to decide whether to include the sentence in the extract or not.

2. Abstract: The abstraction based method relies on the summarizer to form a summary by using only certain components of the original document. A certain freedom exists with the summarizer to form representing sentences using the selected components. A semantic network is build around the document and natural language techniques are used to construct a well represented sentence. The sentence need not necessarily contain the exact words used in the document(s).Abstractive based approaches are very labor intensive and require intensive work on automatic deep understanding of documents.The phases of the abstractive method can be generally classified into three stages.

Stage 1 Content selection sub stage,

Stage 2 Sentence planning sub stage and

Stage 3 Surface generation sub stage. Usually at the surface sub stage, templates will be generated and sentences follow the template. SUMMONS[6] is one the projects that followed the above technique.

3. Capsule: A capsule overview provides an outline of the source text in a semi structural form. The summary will contain the only the key phrases of the document which can be understood only by the knowledgeable user. Boguraev&Kennedy [7] identified the relevant phrasal units and content characterization methods by employing linguistic techniques.

4. Hybrid: This approach combines extraction based techniques with more traditional natural language processing techniques to produce summaries. Firstly the summarizer performs sentence extraction, on which then a process of key concept extraction is performed from these extracted sentences. This is done in two main steps.

Stage 1: First, it reduces sentences by removing any extraneous information where this process is known as sentence compaction. This process uses probabilities learnt from a training corpus.

Stage 2: The reduced sentences are then merged using some rules such as adding extra information about the speakers, merging common sentence elements and adding conjunctives.

The TS can also be classified in terms of input to the summarizer. The single document summarizer abstracts the data of a single text document. The multi document summarizer focuses on multi documents. The latter however is proven to be more useful as the summary occurs from several documents. The similarity of sentences from several documents indicates the importance of the sentence. Hans Peter Luhn[3] describes his research done at IBM, has the most cited paper on text summarization in 1950s. He had initially worked on the frequency of a term. The higher the weightage, the significance of the sentence also arises.

III. METHODOLOGIES

3.1 Feature Selection:

a. Extra weight given to title or heading words of the document. Words mentioned with change in font and cases are considered as relevant keywords.
b. Weight of terms for web engine search applications is dependent on their HTML markup.
c. Term frequency statistics within a category, class or across different categories are used. The terms are then weighed according to relevancy. However the high frequency of a term does not necessarily indicate the importance of the keyword. Also if a keyword is found to be frequent in many documents, it still does not classify the term as relevant. Hence the term document-Inverse Document Frequency is used.

$$Idfi=tf * \log(N/ni)$$

where N denotes the number of documents in the collection, and ni is the number of documents in which term i occurs.

d. Mutual Information (MI) and chi-squared criteria are also used by some summarizers. Therefore for multiple documents

$$MI(t) = \sum_{t \text{ €}\{0,1\}} \sum_{c \text{ €}\{0,1\}} P(t,c) \log P(t,c) / P(t).\, P(c)$$

Where, t→ denotes the term

c→ class label $\begin{cases} 1 \text{ for inclusion in summary} \\ 0 \text{ for not} \end{cases}$

P (t,c) → joint probability of t & c occurring together

P (t) → Probability of document containing term t

P (c) → Probability of document belonging to class c

MI assigns higher relevancy to features that are common in documents considered.

e. Keywords occurring more than a given cutoff are considered a keyword. Title is also taken as a keyword.
f. Thematic word Feature: Maximum Frequency words are called as thematic words. Any word that occurs completely in UPPERCASE is considered as a feature.
g. Usually Nouns are extracted as features. These features can also be selected as a three step process.
1.
1) Morphological Analysis
2) Extraction and Scoring of Noun Phrases
3) Clustering and Scoring of Noun Phrase
Therefore the weight of a term is the sum of all the weights assigned to a term due it font, location, title words etc.

3.2 Sentence Selection

a. Cue Method: Significance of a sentence depends on the occurrence of cue words in the sentence from the selected set of words. Cue words are defined as words that give an important indication of the content of the passage as a whole. Examples of cue words are relevant, important, significant etc.
b. Title Method: All words appearing in the title present in a sentence has its own relevant weights assigned. Sentences that contain title words have a tendency to occur at the beginning and ending of document.
c. Location Method: Studies have shown that paragraphs at the beginning of document have more relevance and information of what's ahead in the document. Hence higher the locality of a

sentence in a document, its weight will be higher. With respect to paragraphs, the sentence points are assigned in descending order from the first line paragraph to the last line of the same.

d. Sentences that contain thematic words and UPPERCASE words are given higher weights.
e. Introductory sentences are also given their due weightage.
f. Proper Noun feature indicates higher weightage for sentence containing proper nouns.
g. Too short or too long sentences are not included in the summary
h. Sentence-to-Centroid Cohesion is another method that is being used. For each sentence s, compute the vector representing the centroid of the document. The centroid is the arithmetic average over the corresponding coordinate values of all the sentences of the document. The similarity is then computed between the centroid and each sentence thus obtaining the raw value of feature for each sentence.
i. Formulating the value of a sentence also includes 1) Sentence length 2) Count of keywords 3) tense 4) type of sentence such as a fact, conjecture or a assertion) 5) Rhetorical relation (such as reason and example) 6) Location of sentence in the document

The rank of sentence depends on the linear sum weightage of similarity of sentence with respect to neighbor sentence, keywords, and presence of title words, capital words and sentence position.

3.3 Basic Algorithm of a Text Summarizer

1. Convert the unstructured text into structured.
2. Stop words are extracted using any algorithm like porter's
3. POS tag is assigned to each word on parsing
4. Store the result.
5. Extract the important key phrases in the text using any feature selection algorithm.
6. The extracted keywords and key phrases are in turn used to rank the sentence.
7. Extract the sentences with the highest rank.
8. If required, repeat process for summary to produce a qualitative summary.

When sentences are selected for the summary, higher probability exists for redundant sentences. The author tends to stress the key topics several times over a document using different sentences but hovering around the same meaning. Therefore sentence similarity function must be run over the initial summarization.

The similarity value is calculated as the vector similarity. If the resemblance value of two sentences is greater than a threshold, the one eliminated is the sentence with low rank based on the features. The sentences are represented as vectors and their similarity values are proportional to the presence of similar words in both sentences.

IV. CHALLENGES OF TEXT SUMMARIZER

1. Adaptation of summary to the level of expertise of user: When a text summarizer is created, the summarized abstract is dependent only on the text. However the summarizer developed must be a function of the user's knowledge too. For example, a doctor is treating a patient for a particular disease. The doctor would like a summarized version of what's new and recent with the disease with respect to the patient. The patient however would like a simplified report without the jargons of medical field involved.
   One method is to keep track of user profile. The jargons will usually be identified as noun phrases. The noun phrases will then be supplemented with a definition when needed so that the resulting summary is less technical and more comprehensible to the user.
2. If summarization is done on an input of several documents, overlapping of themes can occur. The text clustering approach can be implemented here. There will be a cluster of themes. Each cluster represents a theme. The themes are words with high frequency counts. The similarity between a sentence and a theme of a cluster is calculated.
   Another method introduced by Carbonell [8] is by using the MMR measure. The MMR measure penalizes the redundant sentences and rewards the relevant ones.
3. Word Sense disambiguation is one of the commonly faced challenges in the NLP field. When a keyword represents more than one meaning, the challenge here is to know how to resolve the ambiguity with different meanings with respect to a context.
4. How to do we ascertain the goodness of a text summarizer? A good summary must define the question "what" and "to get what". Automatic generated summaries (extracts or abstract) are evaluated mostly intrinsically against human reference or gold-standard summaries (ideal summaries). The problem is to establish what an ideal summary is. Humans know how to sum up the most important information of a text. However, different experts may disagree in considering which information is the best. Evaluation is done with the user abstract against the text summarizer generated report. Recall(R) and Precision (P) are commonly used.

$$P = TSs / (TSs + TSns)$$
$$R = TSs / (TSs + Uns)$$

TSs indicates no of sentences which are present in the text summarizer's abstract but also present in the user abstract. TSns are sentences in the TS's abstract but not in the user's abstract. Uns indicates sentences present in the abstract but not in the abstract of summarizer.In other words, in the case of sentence extraction, the proportion of automatically selected sentences that are also manually assigned sentences is called precision. Recall is the proportion of manually assigned sentences found by the automatic method. The challenge here that the user defined abstract can vary. Due to which the value of precision and recall can vary. The structure of sentence could be different, however semantically mean the same. Precision and recall however, penalizes this. Relative utility [67] has been proposed as a way to address the human variation and semantic equivalence problems in P/R evaluation.

5. Problems of gaps within the summary and dangling anaphors

   As documents contain paragraphs that could contain different topics, there is a possibility of lack of continuation between the sentences. As for dangling anaphors, the values for the frequencies of these tokens do not correctly reflect the occurrence of the concept.

6. Grammatical mistake and plausible output harms the form of resulted summary
7. The source of summarizations can be from a combination of structured, semi structured and unstructured documents.
8. Summaries must be produced fast and in less time with quality preservation.

V. SURVEY OF TEXT SUMMARIZED SYSTEMS

5.1 Extract Based Output

System	Year	Inputs	Features
ERRS [9]	2007	Single & Multi-document	❖ Heuristic- based system has been incorporated ❖ Using the same data structures summaries has been generated
FemSum [10]	2007	Single & Multi- Single	❖ Complex questions have been tried to answer ❖ Using the sentences of syntactic & semantic representation, summaries are produced ❖ Developed using three language independent components: RID, CE, SC
GOFAISUM [11]	2007	Multi-document	❖ Symbolic approach methodology ❖ Techniques involved is tf.idf and syntactic pruning ❖ Summarizes based only on sentences having highest scores.
NETSUM [12]	2007	Single Document	❖ Summaries are generated using machine learning technique ❖ Extraction made based on the best matches of three sentences from the documents
NGD [13]	2009	Multi-document	❖ The number of hits returned by Google is used to work out the semantic distance among concepts ❖ The set of sentences D are clustered into non-overlapping groups of clusters C ❖ Word stemming(Porter's) was used
ILP [14]	2009	Multi-document	❖ This formulation is based on the output of a dependency parser. ❖ summary optimizes an objective function that includes both extraction and compression scores ❖ This system is a valid alternative to existing extraction-based systems
TAC & DUC [15]	2009	Multi-document	❖ Model discusses for sentence selection with a globally optimal solution that also addresses redundancy globally.

			❖ Provides reasonably scalable, efficient solutions for practical problems ❖ It can also be extended to perform sentence compression and sentence selection jointly
NGD[16]	2012	Multi-document	❖ Approach is by optimizing two objectives: Content coverage & Redundancy. ❖ This model applies extraction method and reduces unnecessary information in the summaries.
DESAMC+DocSum [17]	2012	Multi-document	❖ Relevancy, Content coverage, Diversity, Length was taken into account for summarization ❖P-median problem is applied to create the model ❖ To solve the optimization problem a modified DE algorithm is created
VSM [18]	2013	Multi-document	❖ Each unique term represents one dimension in feature vector space. ❖ A subset S D that covers as many conceptual sentences as possible is found ❖ Genetic algorithms using the three operators: crossover, mutation and selection is modeled here

5.2 Abstract Based System

System	Year	Inputs	Features
MultiGen [19]	1999	Multi-document	❖ Similar elements across related text from a set of multiple documents are identified & synthesized ❖ Based on information fusion and ❖ reformulation ❖ Extraction made from sets of similar sentences
Cut & Paste [20]	2001	Single Document	❖ Reduction and sentence combination techniques are used ❖ Identified the key sentences based on techniques like lexical coherence, cue phases & sentence positions.
bi-gram [21]	2008	Multi-document	❖It has used the contextual information to solve feature Sparseness problem. ❖This method is independent on a ❖ kind of language
CLE [22]	2010	Multi-document	❖Maps each data point into r different lines and each map i tries to separate points belonging to class I from others by using label information. ❖CLE becomes more obvious as the percentage of labeled documents or the number of clusters increases
MeSH [23]	2011	Single Document	❖The feature information is enriched by depending on external resources ❖Investigate the different factors influencing the use of citation terms as a means of enriching the representation of a document with

			additional informative synonyms and related terms ❖explore the different factors affecting citation term effectiveness, including section position in the text, and distance from the citation marker, as well as the optimal window size of the citation context boundary
MEDLARS [24]	2013	Single Document	❖goal is to obtain a linear replacement operator RE ❖Rare terms are replaced with linear combinations of their co-occurring terms ❖Dimensionality reduction is employed here

5.3 Hybrid Systems

System	Year	Inputs	Features
Diversity-Based [25]	2010	Multi-document	❖Emphasize on dealing with the text features fairly based on their importance ❖ Two forms based on the structure: first form of the model the diversity rules the behavior of the model, but in second form, the diversity does not rule the model behavior and it works the same in the way as fuzzy swarm-based method
Bi-Gram [26]	2008	Multi-document	❖using contextual information for single- and multi-document summarization ❖to solve feature sparseness problem and the combination method of statistical approaches to improve the performance ❖the proposed method achieved higher performance than other summarization
Zipfian [27]	2006	Multi-document	❖ Index maintenance strategies to be used in on-line index construction for growing text collections ❖ slightly reduced query processing performance
POS [28]	2013	Multi-document	❖ hybrid association rule mining method to identify implicitly ❖ candidate basic rules are reasonable and very common, furthermore, the frequency⁄PMI method can achieve the best performance

VI. CONCLUSIONS

The urgency of information from digital documents and the lack of time have necessitated the need for a text document. The product of the text summarizer must encompass the gist of the text document. The different types of summarizers depending on the kind of summarized output and the input to the Text summarizer were discussed. The techniques for feature and sentence selection are mentioned. The challenges of a text summarizer are also discussed. Certain Extract, Abstract, Capsule and Hybrid examples have been discussed with methodologies employed.

REFERENCES

[1] KarelJezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, UstavInformatiky a softverovehoinzinierstva, 2008.

[2] Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583-598.Oxford University Press, 2005.

[3] H. Luhn. “The Automatic Creation of Literature Abstracts” IBM Journal of Research Development, Volume 2, Issue 2, pp.159-165, 1958

[4] T. Pardo, L. Rino, MNunes: “NeuralSumm: A Connexionist Approach to Automatic Text Summarization”. In: Proceedings of the 4th EncontroNacional de Inteligência Artificial, pp. 210-218, Campinas, Brazil, 2003

[5] Conroy, J. M., Schlesinger, J. D., Goldstein, J. CLASSY Query-Based Multi-Document Summarization. In the Document Understanding Workshop (presented at the HLT/EMNLP Annual Meeting), Vancouver, B.C., Canada, 2005

[6] H. Jing, “Sentence Reduction for Automatic Text Summarization”, In Proceedings of the Sixth Conference on Applied Natural Language Processing, pp.310-315, Seattle, USA, 2000

[7] Boguarev, B., Kennedy, C. Salience based content characterization of text documents. In Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarization, pages 2-9, Madrid,Spain, 1997.

[8] Carbonell, J. and Goldstein, J. (1988). 'The use of MMR, Diversity based Reranking for Reordering Documents and ProducingSummaries', SIGIR 1998, pp. 335-336.

[9] Witte, R., Krestel, R., Bergler, S. Generating Update Summaries for DUC 2007. In the Document Understanding Workshop (presented at the HLT/NAACL), Rochester, New York, USA, 2007

[10] Fuentes, M., Rodr__guez, H., Ferres,D. FEMsum at DUC 2007. In the Document Understanding Workshop (presented at the HLT/NAACL),Rochester, New York USA, 2007

[11] Gotti, F., Lapalme, G., Nerima, L.,et al. GOFAISUM: A Sympolic Summarizer for DUC. In the Document Understanding Workshop (presented at the HLT/NAACL), Rochester, New York USA, 2007

[12] Svore, K., Vanderwende, L., Burges, C.Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 448{457, 2007.

[13] A new sentence similarity measure and sentence based extractive technique for automatic text summarization, Ramiz M. Aliguliyev, Institute of Information Technology of National Academy of Sciences of Azerbaijan, 9, F.Agayev str., AZ1141 Baku, Azerbaijan, Expert Systems with ApplicationsVolume 36, Issue 4, May 2009, Pages 7764–7772

[14] Summarization with a Joint Model,for Sentence Extraction and Compression,Andr´e F. T. Martins_† and Noah A. Smith, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA,Instituto de Telecomunicac¸ ˜oes, Instituto Superior T´ecnico, Lisboa, Portugal, {afm,nasmith}@cs.cmu.edu, NAACL HLT 2009,Integer Linear, Programming for Natural Language Processing,Proceedings of theWorkshop,June 4, 2009, Boulder, Colorado

[15] GenDocSum + MCLR: Generic document summarization based on maximum coverage and less redundancy,Rasim M. Alguliev, Ramiz M. Aliguliyev, Makrufa S. Hajirahimova,Institute of Information Technology, Azerbaijan National Academy of Sciences, 9, B. Vahabzade Street, Baku AZ1141, Azerbaijan, Expert Systems with Applications, Volume 39, Issue 16, 15 November 2012, Pages 12460–12473

[16] DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization, Rasim M. Alguliev, Ramiz M. Aliguliyev, Nijat R. Isazade, Institute of Information Technology, Azerbaijan National Academy of Sciences, 9, B. Vahabzade Street, Baku AZ1141, Azerbaijan, Knowledge-Based Systems 36 (2012) 21–38

[17] Multiple documents summarization based on evolutionary optimization algorithm, Rasim M. Alguliev , Ramiz M. Aliguliyev, Nijat R. Isazade , Institute of Information Technology of Azerbaijan National Academy of Sciences, 9, B. Vahabzade Street, Baku AZ1141, Azerbaijan, Expert Systems with ApplicationsVolume 40, Issue 5, April 2013, Pages 1675–1689

[18] A Scalable Global Model for Summarization Dan Gillick1,2, Benoit Favre21 Computer Science Division, University of California Berkeley, USA , 2 International Computer Science Institute, Berkeley, USA {dgillick,favre}@icsi.berkeley.edu, NAACL HLT 2009,Integer Linear, Programming for Natural Language Processing, Proceedings of theWorkshop,June 4, 2009, Boulder, Colorado

[19] Barzilay, R, McKeown, K, Elhadad, M. Information fusionn in the context of multi-document summarization. In Proceedings of ACL 1999, 1999.

[20] Hongyan, J. Cut-and-paste text summarization. PhD thesis, 2001. Adviser Kathleen R. Mckeown

[21] An effective sentence-extraction technique using contextual information and statistical approaches for text summarization Youngjoong Koa,*, JungyunSeo b a Department of Computer Engineering, Dong-A University, 840 Hadan 2-dong, Saha-gu, Busan 604-714, Republic of Korea b Department of Computer Science and Interdisciplinary Program of Integrated Biotechnology, Sogang University, Sinsu-dong, Mapo-gu, Seoul 121-742, Republic of Korea, Available online at www.sciencedirect.com Pattern Recognition Letters 29 (2008) 1366–1371

[22] Constrained LaplacianEigenmap for dimensionality reduction Chun Chen, LijunZhang, JiajunBu,CanWang , WeiChen Zhejiang KeyLaboratoryofServiceRobot, College of Computer Science,ZhejiangUniversity,Hangzhou310027,China, Neurocomputing 73 (2010) 951–958

[23] Improving MeSH classification of biomedical articles using citation contexts Bader Aljabera, David Martinez a,b, , Nicola Stokes c, James Bailey Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia, NICTA, Victoria Research Laboratory, The University of Melbourne, Victoria 3010, Australia , School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland, Journal of Biomedical Informatics 44 (2011) 881–896

[24] Parallel rare term vector replacement: Fast and effective dimensionality reduction for text T. Berka a, , M. Vajteršica,ba Department of Computer Sciences, University of Salzburg, Salzburg, Austria Department of Informatics, Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovak Republic, J. Parallel DistribedComputing. 73 (2013) 341–351

[25] Fuzzy swarm diversity hybrid model for text summarization Mohammed Salem Binwahlan a,*, NaomieSalim b, LaddaSuanmali c Faculty of Applied Sciences, Hadhramout University of Science & Technology, Yemen, 81310 Skudai, Johor, Malaysia, b Faculty of Computer Science and Information Systems, UniversitiTeknologi Malaysia, 81310 Skudai, Johor, Malaysia, c Faculty of Science and Technology, SuanDusitRajabhat University, 10300 Dusit, Bangkok, Thailand, Information Processing and Management 46 (2010) 571–588

[26] An effective sentence-extraction technique using contextual information and statistical approaches for text summarization Youngjoong Koa,*, JungyunSeo b, Science Direct Pattern Recognition Letters 29 (2008) 1366–1371

[27] Hybrid Index Maintenance for Growing Text Collections Stefan B uttcher Charles L. A. Clarke, Brad Lushman School of Computer Science University of Waterloo,Canadasbuettch,claclark, bmlushma@plg.uwater loo.ca August 6–10,2006, Seattle, Washington, USA

[28] Implicit feature identification via hybrid association rule mining , Wei Wang, HuaXu , Wei Wan State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, Expert Systems with Applications 40 (2013) 3518–3531.