International Journal of Scientific Research in Information Systems and Engineering (IJSRISE)
Volume 1, Issue 2, December-2015. ISSN 2380-8128

Single-Document Summarization Using Latent Semantic Analysis

Oluwajana Dokun, Institute of Graduate Studies and research
Cyprus International University North-Cyprus
E-mail: dokunlewa@yahoo.com
Erbug Celebi, Institute of Graduate Studies and research
Cyprus International University North-Cyprus
E-mail: erbugcelebi@gmail.com

Abstract - *In this study we have evaluated the existing methods of automatic document summarization system and we proposed two approaches in English documents that are based on Latent semantic analysis. Summary selection four existing and two proposed methods for automatic summarization are also used. The evaluated methods that are used include Gong and Liu, Steinberger and Jezek, Murray, Renal & Chaletta, Cross approach and the proposed methods are avesvd and ravesvd. Latent semantic analysis (LSA) is a technique that uses vectorial semantics, for analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA brings out latent relationships within a collection of documents rather than looking at each document isolated from the others. It looks at all the documents as a whole and the terms they contain to identify relationships between them. We have compared the performance of our systems with existing systems in the literature which was developed for this document summarization. The document set used for evaluation of our system is the Document Understanding Conferences (DUC) datasets are document summaries on corpus DUC-2002 and 2004. The evaluation and comparisons of the summaries are performed with ROUGE-L.*

Keywords *– Abstractive Summarization, Extractive Summarization, Information Retrieval, Latent Semantic Analysis.*

1. Introduction

The century we are in today, everything is evolving more than speed of light. The use of internet and documents online has increased exponentially causing a lot of problems for the user to find more precise documents from the millions of documents available. Even the stress of reading the documents and knowing the exact document is a vital problem. This problem can basically be solved using automatic text summarization which is a branch under information retrieval (IR). IR is widely used in search engines, online-book websites, new portals and etc because it makes them more bulky in terms of semantic relationships and context of the documents retrieved. IR is subdivided into many branches one of which is automatic summarization. Automatic text summarization is the creation of reduced type of text by a computer program and the output produce will still contain the most relevant part of the original text and to be more specific automatic text summarization aim at extracting the important sentences from large amount of text in a document and still retain its quality. The goal of this study is to focus on divers ways of automatic text summarization, using singular value decomposition as the algorithm and finding an efficient and effective output method for the summary.

Following sections are organized as follows: Firstly, in section 2 we review previous work of text summarization approaches and evaluation measures. Section 3 explains the main approach used that is LSA approach in details, preprocesses method and step by step of the approach to arrive at the summary of document are explained. Section 4 works on implementation of latent semantic analysis using our proposed system for summarization system. Section 5 also explains the evaluation results of the LSA based single document summarization algorithms using English document sets. Section 6 gives a very brief description of some ideas, concluding remarks and future works.

Corresponding Author
OLUWAJANA DOKUN, Institute of Graduate Studies and research, Cyprus International University North-Cyprus
E-mail: dokunlewa@yahoo.com

2. RELATED WORKS

Researchers have been working actively on text summarization within the Natural Language Processing (NLP) to create a better and more efficient summary. This work started in the late fifties and since then many methods developed from single to multi-document even to multi-lingual text summarization approaches or from extractive to abstractive approach aboveall there have been an increasing in output of summary over the years. Extractive summarization works with the method of finding the salient topics in a text such as Luhn [1] at IBM laboratory, worked on frequency of word in the text. H. P Edmundson [2] used title of the word, cue phase, key method, position method – surface level approach, Daniel Jacob Gillick [3] used classification function to categorize each sentence (sentence extraction) using naïve-Bayes classifier - machine Learning Based Approach. Eduard Hovy and Chin-Yew Lin [4] also, studied on sentence position and later tried to restructure the sentence extraction using decision tree - Statistical Approaches. Gerald Salton [5] worked on automatic indexing which later turned to statistical process that based on term frequency - inverse document frequency algorithm - Graph Based Approaches. Abstractive or non-extractive approach is different from extractive approach but abstractive approach uses extractive approach to generate abstract is that it observes and understand the document, then generates a new summary and this summary does not contain any word from the original document such as Knight and Marcus[6] that used statistical - based summarization to train a system to compress the syntactic parse tree of a sentence in order to produce a shorter but still maximally grammatical version – reduction approach. Daume and Marcus [7] contributed to compression approach using Rhetorical Structure Tree in which they used decision tree to pick the relevant compressed and leave the irrelevant ones – compressive summarization. There are many approaches of text summarization and majority of them are extractive approach because it extracts the important sentences from the input text while text abstraction or non - extractive approach, prove to be the more challenging task, to parse the original text in a deep linguistic way, interpret the text semantically into a formal representation, find new more concise concepts to describe the text and then generate a new shorter text with the same information content. The evaluation of the summaries is another challenging part of document summarization because there is still no proper or ideal summary for document but different evaluation approaches have been used for text summarization in general.

3. LATENT SEMANTIC ANALYSIS

Latent Semantic Analysis (LSA) is the combination of algebraic and statistical methods and this technique brings out the hidden structure of words, between words, sentences or document. The main ideas of LSA is that it extracts the input document and convert to sentence – term matrix and process it through an algorithm called singular value decomposition(SVD). The purpose of the SVD is to find relationship between word and sentences, reduce noise and also model the relationship among sentences and words. Finally, output is obtained from SVD algorithm. LSA main algorithm to text summarization is divided into three steps: creation of sentence - term matrix, applying SVD to matrix and selection the sentence for the summary.

In this section, we firstly describe creation of sentence-term matrix. Secondly, applying SVD to matrix and different algorithms for selecting sentence in latent semantic analysis

3.1 Creation of sentence - term matrix

In LSA, creation of sentences by term matrix is based on vector-space model (VSM) that is the arrangement of bags of words into their sentence by term matrix. Matrix is the representation of data into rows and columns; where rows represent the words, columns represent sentences and each data is filled into their cell. The creation of matrix is a very difficult task in LSA because it must pass through a pre-processing method before it becomes full sentence – term matrix

3.2 Applying SVD to matrix

SVD is based on a theorem from linear algebra in which a rectangular matrix A is decomposed into three matrices - an orthogonal matrix U, a diagonal matrix $\sum$, and the transpose of an orthogonal matrix V. The purpose of SVD is using a dimensional matrix set of data points and reducing it to a lower dimensional space. SVD is used to reduce dimension of term-by-document matrix. This technique

also reveals the latent data while removing the noise. The computation of the

SVD as follows:

SVD decomposes a matrix (A) into three matrixes.

$$A = U\Sigma V^{T}$$

where, U is a matrix that their columns are the eigenvectors of the *AAT* matrix. This matrix is called left eigenvectors and it represents concept-by-term relation.

Σ is a matrix that their diagonal elements are the singular values of *A*. Its non-diagonal elements are 0. The matrix represents concept-by-concept relation.

V is a matrix that their columns are the eigenvectors of the *ATA* matrix. This matrix is called right eigenvectors and it represents concept-by-document relation.

V^{T} is the transpose of V.

From reduced dimension, a suitable k (rank approximation value) value should be chosen to reduce the dimension of the LSA space. The amount of dimension reduction is determined by value of k so it does not matter whether k is large or small but just to reduce unimportant details and to yield good retrieval result.

$$A = U_o \Sigma_o V^{T}_o$$

3.3 Sentence Selection

Sentence selection is done after creating the input matrix and singular value decomposition of the matrix. The next step is ranking of the sentences based on the scores and selecting the ranked sentences based on the type of algorithm that is used for document summarization. In this study four of them will be explained in details and their various algorithms used LSA for document summarization.

3.3.1 Different algorithms for selecting sentence in latent semantic analysis

Gong and Liu [8] pioneers started the use of LSA for text summarization. They started by creating the term by sentences matrix They started by creating the term by sentences matrix and their reasons of SVD to matrix is of two views: from transformation aspect that is it gives a mapping between the m-dimensional space spanned by the weighted term-frequency vectors and the r-dimensional singular vector space with all of its axes linearly-independent. From semantic point of view, the SVD obtains the latent semantic analysis from the document represented by matrix that is the breakdown of the original document into r linearly-independent base vectors or concepts. After performing the SVD on term sentence matrix, a singular value matrix and the right singular vector matrix V^{T}. In the singular vector space, each sentence is represented by the column vector of V^{T} and then picks the p^{th} right singular vector from matrix V^{T} which means that is selecting the sentence which has the largest index value with the p^{th} right singular vector, and included it to the summary. Finally, until p reaches the predefined number that is being defined by the user it then, the operation will terminate else increment p increases by one, and go to back again

The main disadvantages of Gong and Liu's method is that when sentences are extracted the top topics are treated as the same as equally concepts. Secondly, the related only one sentence from each concept showing that the same number of sentence collected is the same as the dimension and the larger the sentence the less the important concept is picked. Steinberger and Jezek [9] approach selects sentences through vectorial representation into matrix that has the highest length using sentence vector

$$L = \sqrt{\sum_{j=1}^{n} V_{ij} \times \Sigma_{jj}}$$

where *L* represents length of the score, *V* is a matrix that their columns are the eigenvectors of the *ATA* matrix and Σ is a matrix that their diagonal elements are the singular values of *A*.

Murray et al. [10] in their approach more than one sentence can be collected from the topmost important concepts, placed in the first rows of the matrix rather than using extracting the best sentence for each topic. Decision of how many sentences will be collected from each concept is made by using matrix. The value is decided by getting percentage of the related singular value over the sum of all singular values, for each concept. Murray et al. approach solves the problems of Gong & Liu's approach of selecting single sentence from each concept, more than one sentence can be chosen even they do not have the high-

est cell value in the row of the related concept. And also, the reduced dimension makes it different from other approaches.

Makbule Gulcin Ozsoy [11] used Cross method to improvise on Steinberger and Jezek approach. In this approach input matrix creation and SVD calculation steps are executed as in other approaches and then the matrix is used for sentence selection purposes but between the SVD calculation step and the sentence selection step, a pre-processing step is placed and the purpose of step is to remove overall effect of sentences that are not related to the concept, leaving only the most relevant sentences related to that concept. Mathematically, for each concept, that represents rows of the matrix, the average sentence score is calculated. Then the cell values which are less than or equal to the average score are set to zero. After that Steinberger and Jezek approach is followed with little modification that is by adding up the concept scores with values after the preprocessing step

4. IMPLEMENTATION

4.1 Introduction

We developed an application to provide a strong single document automatic summarization system. The system has been implemented in the Java programming language and JAMA library [12] was used as part of the application. JAMA is a basic linear algebra package for Java. It provides user-level classes for constructing and manipulating real, dense matrices and also consists of pairs or triples of matrices, permutation vectors, and the like, to produce relevant and accurate results.

4.2 Implementation Processes

Step 1: Loading is a primary step of inputting the text into the buffer, and make some processes for summarization. Our system load the folder file into the buffer, checking the input document if it available or not, changing the text into the lower case state and also differentiate between similar words. All these are to improve the accuracy and eliminate redundancy in the system.

Step 2: Tokenization is converting a stream of characters into a stream of processing units called tokens. After the system has placed the text inside buffer, it splits the words from sentence into their units and then remove the punctuation marks, parenthesis, quotes, whitespace positioning, etc. so that sequence of tokens can be obtained.

Step 3: The next step is stemming, stemming the process of reducing different forms of a word into its root form and the purpose for using stemming is to reduce memory usage for storing the words. In our system Porter's stemming algorithm which is the best known stemmer is used and it helps to remove prefixes and suffixes as well as some transformation rules.

Step 4: After tokenization and stemming we still discover that most sentences use period at the end of each sentences such Mrs. Prof., so we have to follow the rules of sentence discrimination to solve this problem such as rule like not to break a sentence when the sentence contains the numeric words in it.

Step 5: Term frequency is a mathematical matrix that fills the cell with frequency of term that occurs in a collection of document. In our system each row represents the document in the collection and the column represent the term in the document. The higher the terms and the frequency, the greater the numbers of times in which the word occurs in the document.

Step 6: Sentence selection using LSI, we developed two systems that use the above steps up to creation of SVD matrix. After creating the matrix, sentence selection is the next steps which were also done by other approaches as seen in section four and this approach of select their summary is from V^T of the SVD matrices. Gong and Liu proposed that row order indicates the importance of the concepts such that the first row represents the most important concept extracted. In this study, we assume that instead of using Gong and Liu approach of selecting the highest from each concept of the entire row then we find the average of the entire concept of each sentences and one sentence is picked from the average most important sentences of each concept, and then second sentence is chosen from the second average most important concept of each concept; and this process continues until all predefined number of sentences are collected

4.3. The Proposed Method avesvd

The avesvd method is an extension to the Gong and Liu approach. avesvd uses the above steps up to the creation of SVD matrix. After these steps, sentence selection is the next step which was also done by other approaches as seen in chapter Four. avesvd selects its sum-

mary from V^T of the SVD matrices. In Gong and Liu approach, they proposed that row order indicates the importance of the concepts such that the first row represents the most important concept extracted instead of using Gong and Liu approach to select the highest from each concept of the entire row then we find the average of the entire concept of each sentences column using equation 1 below to calculate the average

$$k = \frac{1}{n}\sum_{j=1}^{n} a_j = \left(\frac{a_1 + a_2 \ldots\ldots a_n}{n}\right) \qquad (1)$$

where each column is denoted by a_j and n is total number of column

Finally, one sentence is selected from the average most important sentences of all the concepts. Then second sentence is chosen from the second most important sentence and this process continues until all predefined number of sentences are collected as shown in Figure 1.

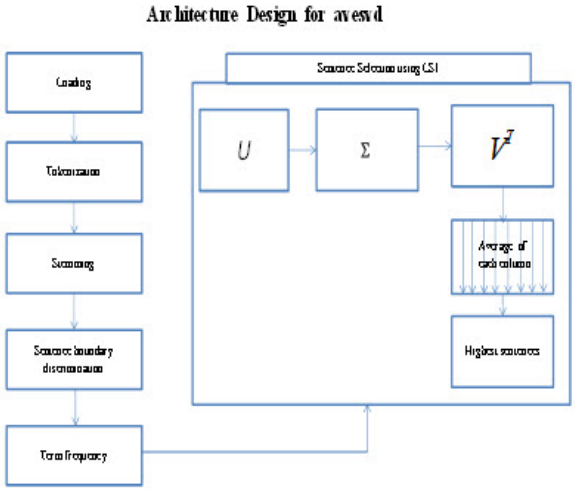

Figure 1-showing the architectural design of avesvd

4.4. The Proposed Method ravesvd

The ravesvd method is similar to the avesvd approach following the steps from Figure 2 approaches based on Latent Semantic Analysis. In this step first the input documents are represented in a matrix form, using the example above as the input file and then SVD calculation is done. After these steps, the system selects its summary from the average of all columns from the V^T matrix in the SVD selecting the sentences for the summary using the equation 2 below.

$$k = \frac{1}{n}\sum_{j=1}^{n} a_j = \left(\frac{a_1 + a_2}{n}\right) \qquad (2)$$

where each column is denoted by a_j and n is total number of column.

The system calculates the average of the entire reduced concept of each sentences column from the V^T matrix and first sentence is chosen from the highest average score of reduced concepts, and then second sentence is chosen from the second average most important score of reduced concept; and this process continues until all predefined number of sentences are collected.

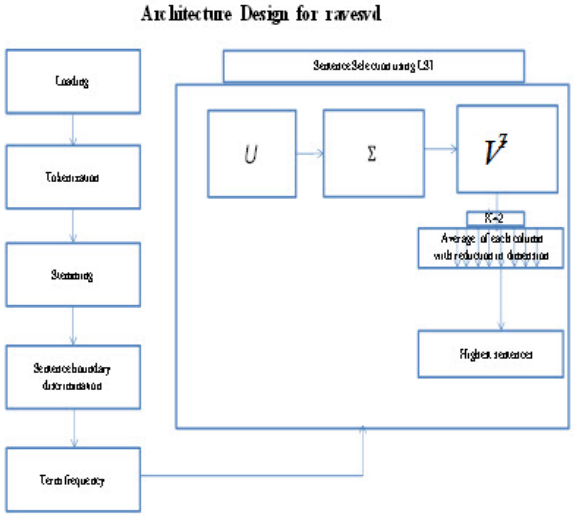

Figure 2 - The architectural design of ravesvd

The idea behind this ravesvd to reduce the dimension of the concept (k = 2) and not to lose any concept that is available after the reduction of dimension. This method of reducing can cause loss of many topics but if not we might include less important topic and noisy in the summarization.

5. Experimental results

Four tests experiment with our systems for document summaries on corpus DUC-2002 and 2004. The DUC 2002 and DUC 2004 document sets consist of 482 and 567 document sets respectively. Out of the 567 documents present in the DUC-2002 document set, we were left with 556 documents and out of 482 documents in DUC-2004, 475 were used after the clean-up processes. This is done with English dataset for comparing the results of these different approaches of text summarization methods using ROUGE as standard evaluator. In this section, we worked on two different dataset; preparation for ROUGE setup and the

result from the proposed methods mentioned in section 4 were also examined. For each experiment the corresponding proposed method were applied and the composed summaries were evaluated. The obtained results are presented and discussed.

5.1 Evaluation Setup

Our evaluation is based on three evaluation steps: Latent Semantic Analysis approach, Output part and ROUGE evaluation part

5.2 Latent Semantic Analysis approach

The system generates latent semantic analysis for four existing systems and two proposed systems were used as an input document given by DUC as peer summaries. The input documents used JAVA and JAMA as part of the library for the process of the input and an XML file is produced as output. XMLs for all the documents in the DUC document set were retrieved and stored for use by all summarization system.

5.3 Output part

The second part of the XML is producing the model summaries which is reference summaries that are also in XML file format. In essence, the output comprises both model and peer summaries as XML file. Finally, both the model and peer summaries extracted must have a corresponding summarizer ID number created for ROUGE evaluation step–up.

5.4 ROUGE Evaluation component

The ROUGE system requires as input an XML file that specifies the peer summaries system summaries and model summaries all these have to be evaluated together. Figure 9 is an example of XML file format require for ROUGE evaluation. This system was written in order to accept an XML file format as input for ROUGE evaluation. While running ROUGE, several options can be chosen some of which specify preprocessing tasks but we produce for ROUGE evaluation as shown in Illustration above. The performance of the evaluation of the text summarization algorithms were described in this study and we obtain ROUGE-L results using Perl ROUGE-1.5.5.pl -a ./ROUGE_EVAL_HOME/thesis.xml for ROUGE-L as our results metrics and discussions are made based on the outputs.

5.5 Results

The ROUGE toolkit was run to evaluate the summaries metrics recall, precision, and F-measure. The 2002 and 2004 datasets in English are datasets for the evaluation of the summarization systems. In order to compare LSA based approaches with other approaches we used Duc2002 and DUC2004 datasets, different resources are used and their evaluation results for summarization are collected. The tables below indicate the ROUGE-L, scores obtained from running the ROUGE evaluation toolkit to compare the most important sentence chosen by the three model summaries and the summary generated.

Table 1 - ROUGE-L Scores for 2002 Dataset

	ROUGE_L RESULT FOR 2002 DATASET		
APPROACHES	RECALL	PRECISION	F-MEASURE
Gong And Liu	0.2671	0.2564	**0.2531**
Murray	**0.2967**	0.2306	0.2507
Cross	0.1179	0.2879	0.164
avesvd	0.2168	0.2474	0.2269
ravesvd	0.2470	0.2392	0.2397

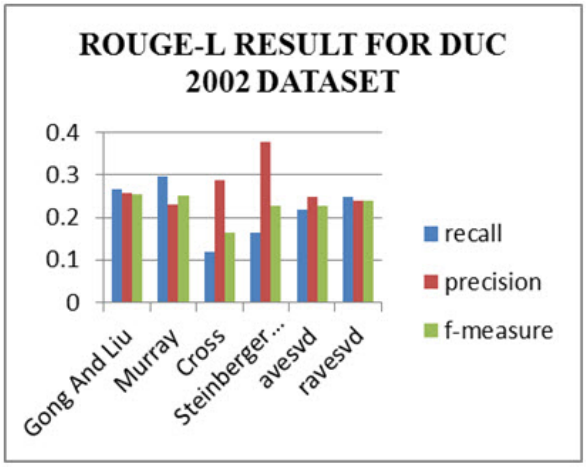

Figure 3- ROUGE-L scores for 2002 Dataset

Table 2 - ROUGE-L scores for 2004 Dataset

	ROUGE-L RESULT FOR 2004 DATASET		
APPROACHES	RECALL	PRECISION	F-MEASURE
Gong And Liu	0.6517	0.1715	0.2619
Murray	**0.6536**	0.1394	0.2186
Cross	0.5962	0.4624	0.5142
Steinberger And Jerek	0.6473	**0.4708**	**0.5374**
avesvd	0.6346	0.1971	0.2993
ravesvd	0.6375	0.1666	0.2629

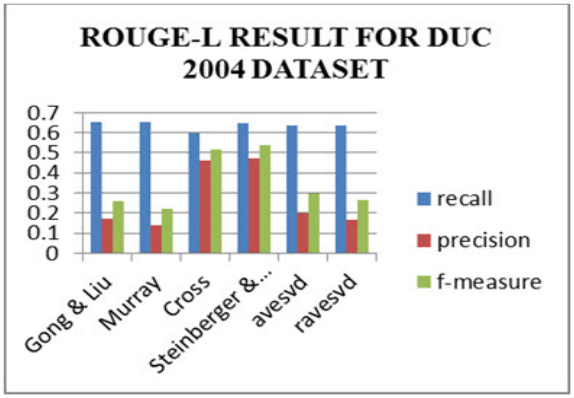

Figure 4- ROUGE-L scores for 2004 Dataset

From the Table 1 and Table 2, it has been observed that Murray method has the highest score in recall both DUC-2002 and DUC-2004 datasets, Steinberger & Jezek has the highest score in precision for both DUC-2002 and DUC-2004 datasets, Gong and Liu has the highest score in F-measure DUC-2002 dataset, Steinberger & Jezek has the highest score in F-measure for DUC-2004 dataset, ravesvd became third in both recall and F-measure DUC-2002 and ravesvd became fourth in both recall and F-measure DUC-2004. Overall, we believe that the results are encouraging but still need more improvement to achieve better results.

. 6. Conclusion

The system we have built is a LSA-based summarization system for single document summarization system. LSA presumes that words that are close in meaning will occur in similar pieces of text and analyzes relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. In this study, we have explored some relevant linear algebra concepts, attempted to automatically summarize text using the LSA approach and we evaluated generated summary using the ROUGE evaluation toolkit. To speed up process of calculating SVD, we also took advantage of predefined library called JAMA library. The results show that both systems extracted sentences reasonably well from the original text. We chose to select a sentence with the largest value after finding the average of each in each column of V^T for each topic, without assuming that any concept of the sentence is less important. The disadvantages are sentences might have equal average scores, but how to prioritize them is a problem. Therefore, sentence selection should be further refined and from the result we observed that with reduction in dimension it only have little effect on the results.

7. REFERENCES

H.P.Luhn, "The Automatic Creation of Literature Abstracts" IBM journal April, 1958.

H. P. Edmundson "New Methods in Automatic Extracting" Journal of the Association for Computing Machinery, Vol. 16, No. 2, April 1969.

Daniel Jacob Gillick "The Elements of Automatic Summarization" Electrical Engineering and Computer Sciences, University of California, May 2011.

Eduard Hovy and Chin-Yew Lin, "automated text summarization and the SUMMARIST system" Information Sciences Institute of the University of Southern California, 1998.

Gerald Salton and Christ Buckley" Term Weighting Approaches in automatic Text Retrieval", Department of Computer Science, Cornell University, New York, November 1987.

Kelvin Knight and Marcu Daniel "Statistical-Based Summarization One step: Sentence compression" Information sciences Institute and Department of Computer Sciences University of Southern California, 2002.

Hal Daume III and Daniel Marcu "A Tree-Position Kernel for Document Compression Proceedings of the Document Understanding Conference", Boston, MA. May 6-7, 2004.

Yihong Gong and Xin Liu "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis", New Orleans, Louisiana, C & C Research Laboratories, USA. SIGIR'01, September 9-12, 2001.

Jezek and Steinberger "Automatic Text Summarization" The State of Art 2008 and the challenges, Bratislava. 2008.

Gabriel Murray, Steve Renals, Jean Carletta "Extractive Summarization of Meeting Recordings" Centre

for Speech Technology Research, University of Edinburgh, Scotland, 2005.

Makbule Gülçin Özsoy "Text Summarization Using Latent Semantic Analysis" Graduate School of Natural and Applied Sciences, Middle East Technical University, Ankara.2011.

http://math.nist.gov/javanumerics/jama/[last reviewed 9, Feb 2011].