Eng.& Tech. Journal ,Vol 30 , No.5 , 2012

Personal Text Summarization in Mobile Device

Dr.Alaa Kadhim
Computer Science Department, University of Technology/Baghdad
Email:Dralaa-Student@Yahoo.Com

Received on: 7/7/2011 & Accepted on: 5/1/2012

ABSTRACT

This paper presents a hybrid text summarization for mobile device to summarize a selected text. The system can be proceeds by statistic or heuristic methods. With the statistic and heuristic the summary is found based on combined statistic features and heuristic features like word frequency, position, length of sentences, and similarity with the document title. The results shows that the time with proposed system is less than without it during the retrieving the text with selected keywords.

Keywords: Text summarization, mobile device, statistical text features, heuristic text features

التلخيص الشخصي للنصوص في أجهزة الهواتف النقالة

الخلاصة

هذا البحث يقدم نموذج جديد لنظام مهجن لغرض تلخيص النص في أجهزة الهواتف النقالة. النظام يعالج النص بطرق إحصائية أو الإحصائية موجهة . من خلال الطرق الإحصائية يتم تلخيص النص بالاعتماد على خصائص إحصائية، والأسلوب الأخر أي الإحصائية الموجهة يتم تلخيص النص بالاعتماد على الخصائص الإحصائية والموجهة مثل تردد الكلمة، مواقع وأطوال الجمل وتماثل العنوان للنص ككل. النظام خرج بمجموعة تمثل أفضل نسب لمعاملات الخصائص المستخدمة في أيجاد تلخيص النص بأفضل صورة بالاعتماد على الكلمات المختارة.

INTRODUCTION

Advance of mobile network creates business opportunities and provides value-added service to user access to the internet through mobile phone should be extend to decision support in an organization . summarization techniques have been applied to summarize data in mobile phone**[1].**

With the increase of textual information, summarizing document is becoming an important issue. Text summaries allow users to rapidly consult retrieved documents and decide on their relevance **[2].**

Automatic text processing is a research field that is currently extremely active; one important task in this field is automatic text summarization, which consists of reducing the size of a text while preserving its information content **[3].** Summarization can be defined as the selection of a subset of the document sentences which is representative

https://doi.org/10.30684/etj.30.5.7
2412-0758/University of Technology-Iraq, Baghdad, Iraq

of its content. This is typically one by ranking the document sentences and selecting those with higher score and with a minimum overlap **[2]**, in general there are two types of automatic text summarization which they are extract and abstract **[4].**

Extraction-based summarization uses statistical basis or heuristic method or a combination of both to extracts important sentences from an article by statistically weighting the sentences or heuristically such as position information or title similarity for scoring sentences **[4],** it can operates in two modes: generic summarization, which consists in abstracting the main ideas of a whole document and query-based summarization, which aims at abstracting the information relevant for a given query **[2].** While abstract summarization involves rewriting the original text in a shorter version by replacing wordy concepts with shorter ones **[4].**

SUMMARIZATION TYPES

The core formula for text summarization is as simple as to select sentences with special characteristics and put these together in a summary. There are two major types of text summary: Abstract and Extract.

a) Extract Summarization

The core formula for extraction-based summarization is as simple as to select sentences with special characteristics and put these together in a summary .[5] The summarized text is extracted from the original text on a statistical basis or by using heuristic methods or a combination of both. With the sentence extraction, extracts important sentences from an article by statistically weighting the sentences. Or heuristics such as position information are also used for summarization. For example, summarization system may extracts the sentences which follow the key phrase "in conclusion", after which typically lay the main points of the document [6]. This means that the extracted parts are not syntactically or content wise altered [5].

b) Abstract Summarization

An abstract is a summary, a least some materials of which does not exist in the original document (e.g. point of view on the document, paraphrase, etc.). Or it is an interpretation of the original text, which assume semantics level of representation of the original text and involve linguistic processing at some level [7]. The process of producing it involves rewriting the original text in a shorter version by replacing wordy concepts with shorter ones, which requires the ability to manage various hard AI problems. For example, the phrase "He ate banana, orange and pear" can be summarized as "He ate fruit" [5].

FEATURES FOR TEXT SUMMARIZATION

Some features are useful in calculating the importance of a sentence within document, which often increase the candidacy of a sentence for inclusion in summary [8], these features are of some kinds, like statistical based on the frequency of some elements in the text, linguistic extracted from a simplified argumentative structure of the text; or heuristic based on sentence length or position and some other features, as describes bellow:

STATISTIC

Based on the frequency of some elements in the text; which give different information about the relevance of sentences for the summary. These features are sufficiently relevant for the single document summarization task [9], some of those features are listed below:

- **a-** Numerical data: Sentences containing numerical data are scored higher than ones without numerical values. A constant value is added to the score of the line (Default value '1') [8].
- **b-** Word Frequency (WF): Open class words (content words) which are frequent in the text are more important than the less frequent [8].

$WF(w)= \sum_i^{w€d} wi$

w is the word, d is document

LINGUISTIC

Extract from a simplified argumentative structure of the text, which assume semantics level of representation of the original text and involve linguistic processing at some level [7].

HEURISTIC

It is based on sentence length or position and some other features as follows:

- **a.** Position Score: The assumption is that certain genres put important sentences in fixed positions. For example, newspaper articles have the most important terms in the first four paragraphs.
- **b.** Title: Words in the title and in following sentences are important and get high score [2].

$TF(w)= \sum_i^{w€t} wi$

TF(w) is the number of time w occurs in the title

- **c.** Similarities between sentences in an article [7].
- **d.** Indicative Phrases: Sentences containing key phrases like "this report …"[8].
- **e.** Sentence Length: The score assigned to a sentence reflects the length of the sentence, normalized by the length of the longest sentence in the text [8].

CELLULAR MOBILE TELEPHONY

Cellular mobile telephony is the popular example of mobile communication systems. The cellular mobile phone system is characterized as a system ensuring bidirectional wireless communication with mobile stations moving even at high speed in a large area covered by a system of base stations.

The cellular mobile system can cover whole countries. Moreover, a family of systems of the same kind can cover the area of many countries. Initially, the main task of a cellular system was to ensure the connections with vehicles moving within a city and along highways. The power used by cellular mobile

stations is higher than that used by the wireless telephony and reaches the values of single watts .

PROPOSED SYSTEM

The proposed system contains three main steps, keywords extraction, sentence ranking selection and text summarization. The below block diagram shows that.

First Stage : Keywords Extraction

In this stage the classical keywords extraction steps have been used as in the following points:

Ignore stop words

Stopword referred to counter the obvious fact that many of the words contained in a document do not contribute particularly to the description of the documents content; they are very frequent and non-relevant words. For instance, words like "the", "is" and "and" contribute very little to this description. Therefore it is common to remove these so-called stop words prior to the construction of the document descriptions, leaving only the content bearing words in the text during processing.

Stemming

Removing suffixes by automatic means is an operation which is especially useful in the field of information retrieval. Terms with a common stem will usually have similar meanings, for example:

CONNECT – CONNECTED – CONNECTING -CONNECTION

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION to leave the single term **CONNECT**. The proposed system uses a porter algorithm, and this section shows how the system performs stemming. It is removes about 60 different suffixes, which involves a multi-step process that successively removes short suffixes, rather than removing in a single step the longest suffix.

Word Ranking

In this section the score or weight of all words in the document by three features are finding, to use them for sentences weighting, as follows

Feature one: WF/ID of each word w

Statistically find the weight of words by counts the appearance of them in the whole document.

$$\mathbf{WF(w)= \sum_i^{w \epsilon d} wi \quad ....(1)}$$

w is the word, **d** is document
For each word (**w**) in the document count all appearance of it in the document.
Feature two: Similarity with the title of document of each word w, Heuristically for each word in document check if it is appears in the document title or not.
TF(w)= 1 if w appears in the title, otherwise 0
Feature three: WS of each word w**,** Statistically for word (w) find the weight of it by counts the number of sentences it will appears in it.
$S(w)=\sum_i^{s\in d} w\in s_i$
S(w) is the number of sentences in which **w** occurred, **s** is sentence, **d** is document.
For each word (w) in the document count the number of sentences that it is appears in it.
Sentence ranking
The sentences need to be scored, by using weights of its words. It depends on the methods that are to be used in the system, which uses the specified features far each using methods. The proposed system uses two methods which are statistics and heuristics, for each of them used the three specified features as follows:
1- Statistics: Using simple statistics, to determine the summary sentence of a text by using the words weight; and three statistics features:
Feature One: Weight Sentence WS

$$WS(s)= \sum_i^{w\in s} w_i \qquad \ldots(2)$$

For each sentence (**s**) in the document count all its words.
Feature Two: Sentence Frequency SF of each sentence S

$$SF(s)=\sum_i^{w\in S} WF(w) \qquad \ldots(3)$$

Where **S** is the sentence, **WF(w)** is the frequency of word **w** that accurse in the sentence **S.**
Feature Three: TF/ISF the sentence score is the summation of its words scores or weights. The score of each word w is given by the following formula:

$$TF/ISF=F(W)*(\log n/ S(W)) \qquad \ldots(4)$$

Where **F(w)** is the frequency of **w** in the sentence, **n** is the number of words in the sentence and **S(w)** is the number of sentences in which **w** occurred.

Heuristics: Using heuristic features, to determine the summary sentence of a text, the system used also three features which are:

Feature first: Similarity with the title, specify if the words in the sentence occurs in the document's title or not. Give higher degree to sentence's that have more words similarity with words in the document's title.

$$TF(s)= \sum_i^{w \in s} TF(wi) \quad \ldots(5)$$

TF(s) is the number of time w occurs in the title

Feature two: Position of the sentences, the assumption is that certain genres put important sentences in fixed positions; the proposed system assumes that the important sentences are those that are in the first three sentences in document.

PF(s)= 1 if s is one of the first three sentences, otherwise 0

Feature third: Length of the sentences, the score assigned to a sentence reflects the length of the sentence as to prevent long sentence from getting higher score, and prevent ignoring the small sentences. The proposed system assigns score of sentence depending on the length of it, to normalized length by the longest sentence in the document, as appears in this formula:

Word-count = number of words in the text.
Worde-count = number of words in the text.
Average sentence length (ASL) = Word-count / Worde-count
W_{sl} = (ASL * Sentence Score)/ (nr of words in the current sentence

Sentence selection

The proposed system uses statistics and heuristic methods for finding summary, after finding score of each sentence by the specified methods and features the combination function is required to ranking sentences with different weights for giving them the final sentence score.

THE IMPLEMENTATION

In this work, Visual Basic .Net (2008) in Smart Phone was used to program the proposed text summarization. The proposed application has multi-pages in the GUI. The programming language is suitable for Smart Phone Platform with the compatibilities of their devices and applications.

Figure 2 illustrates the first page which is represent the loading page of the proposed work.

Figure 3, illustrates the mobile page that contains the selected text that the proposed work will summarize it.
Figure 4, shows the extracted keywords from the selected text, the have 2 option, the first use the default and in the second the user must select at least one extracted keyword.
Figure 5 illustrates the summarized text from the original text with user selected keywords.

CONCLUSIONS

The mobile devices have special hardware capabilities which are different from the PC, therefore, the applications programming must be compatible with these devices. The text summarization programming in mobile devices different from the PC in several states. There are several conclusions points from the implementation of the proposed work which are:

1- The proposed work enhance the performance and options of mobile device text summarization facilities.
2- This work decrease the time of text summarization operation with the option of selected keywords.
3- The implementation of work provide a good facility for text retrieval and summarization in the mobile device, also it present a good text services in the future

REFERENCES

[1] Internet Page '**An information delivery system with automatic summarization for mobile commerce**',C.Yang,2005. *www.elsevier.com/locate/dss*

[2] Massih-R. , Patrick G., **"Self-Supervised Learning for Automatic Text Summarization by Text-span Extraction"**, Proceedings of the 24th Eu ropean Conference on Information Retrieval (ECIR), pp. 55-63, Pisa, Italy, 2001.

[3] Mehmed, K.**,** "**Data Mining—Concepts, Models, Methods, and Algorithms**"**,** Wiley & Sons IEEE Press, 343 pp ISBN 0-471-22852-4, volume 2, Number 3, Totowa, USA, 2003.

[4] Naoaki, O., Yutaka M., Naohiro M., Hironori Tomobe, and Mitsuru I., **"Extracting Characteristic Sentences from Related Documents"**, Proc. 6th Int'l Conf. on Knowledge-based Intelligent Information Engineering Systems & Applied Technologies(KES2002), IOS Press/Ohmsha (ISSN:0922-6389), pp. 1257--1261, Crema, Italy, September 2002.

[5] Delort, J., Bouchon B. and Rifqi M., **"Enhanced Web Document Summarization Using Hyperlinks"**, Proc. 14th ACM Conference on Hypertext and Hypermedia (HT'03), pp.208–215, United Kingdom, 2003.

[6] Weiguo, F., Linda W., Stephanie R., and Zhongju Z., **"Tapping into the Power of Text Mining"**, Communications of ACM, Volume 49, Issue 9, pp. 76 – 82, New York, USA, 2006.

[7] Joel, L., Alex A., Celso A., **"Automatic Text Summarization using a Machine Learning Approach"**, paper presented at the SBIA '02: sixteenth Brazilian Symposium on Artifial Intelligence, porto de Galnhas/ Recife, Brazil, 2002.

[8] Nima, M., **"A Persian text summarizer"**, Master Thesis, Department of Linguistics,
Stockholm University, Sweden, January 2004.

[9] Massih, R., Nicolas U., and Patrick G.,**" Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms"**, ECIR, LNCS 3408, Springer-Verlag Berlin Heidelberg, USA, 2005

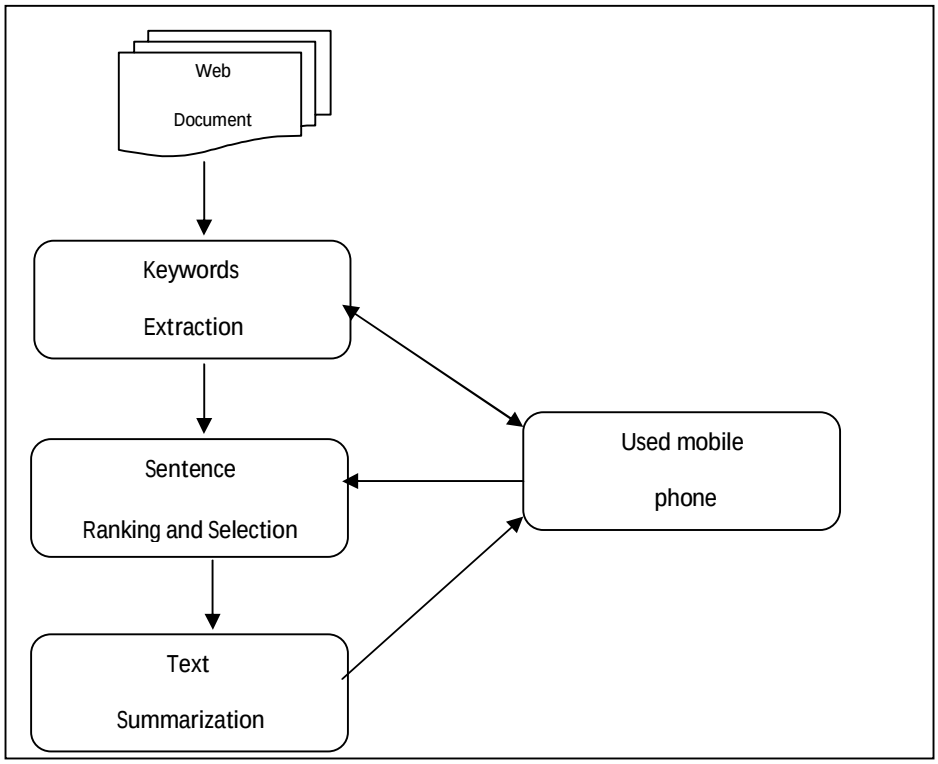

Figure (1) proposed system

Figure (2) start page in mobile device emulator

Figure (3) text before summarization in mobile device emulator

Figure (4) choose select word or default option and ratio

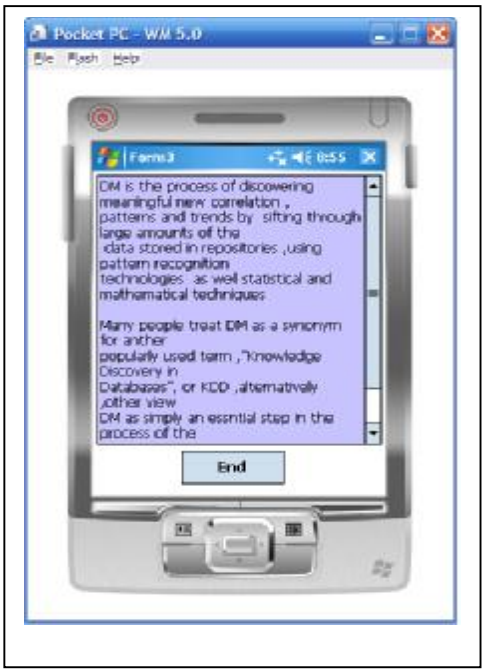

Figure (5) text after summarization