Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEARCH IN TEXT DOCUMENTS BASED ON N-GRAMS AND FOURIER WINDOW TRANSFORMATION

K.YA KUDRYAVTSEV
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoye highway 31, 115409, Moscow, Russian Federation
E-mail: const58@mail.ru

ABSTRACT

This paper is of theoretical nature and is devoted to the development and theoretical justification of the new Full Text Search method in documents. The idea of the method is to partition the text into N-grams, to construct "spectrograms" of a text document using the windowed Fourier transform. Then a fast Fourier transform algorithm is used and "spectrogram" of the text of the document is compared to the "spectrogram" of the search line in special "control" points. This paper provides a rigorous mathematical justification of the proposed method, the possibility of forming conclusions on the presence of the search line in the document is proven. A generalized algorithm of a new method of full-text search is presented.

Keywords: *Full-text search in databases, Window Fourier transform, Fast Fourier Transform, Spectrogram text, N-Grams.*

1. INTRODUCTION

One of the most common ways of finding relevant information in a text document is a method in which the search is performed by using so-called N-grams, which are determined by dividing a search string (or phrase) into a set of substrings (or N-grams) of fixed length [1]. A text document is also represented as a set of N-grams, and it is necessary to determine whether the N-grams contained within the search string match those of the text document, and if so, in what proportion. Instead of N-grams, single words can be used; in this case doing so would determine whether a word from the search string features in the text document, and if so, in what proportion.

Differently structured indices are built up to work with N-grams, such as inverted index, B-trees or suffix trees [2],[3],[4],[5]. The advantage of using an N-gram index is the language neutrality that doing so allows, as well as the error tolerance in spelling [6],[7],[8],[9].

An N-gram index can be very large and require a lot of disk space, which leads to a decline in performance in search queries [10]. Compression techniques are used to reduce the size of the indices [11], or individual non-overlapping N-grams (or, as a rule, words) can be selected [12]. The N-gram set of keywords, which is formed at an advanced stage of the document processing, can be replaced. However, the absence of some keywords will exclude the document from the search list.

Therefore it is desirable that such a search system would not require the use of a bulky index structure to store N-grams, and that it would include the preliminary introduction of a keywords list.

N-grams comparing is done symbol by symbol, which leads to significant time losses. Moreover, spelling mistakes (misprints) in the text do not allow choosing relevant documents as the text in N-grams and in search line will not match.

This is why there is preference for another way of comparing of N-grams other than symbol by symbol one.

This paper offers a fundamentally new original method for determining compliance of a search line to text lines in a document. The method is based on the use of N-grams, the theory of spectral analysis, the windowed Fourier transform and the fast Fourier transform algorithm. The principal difference is that the text is seen as time series and is converted to a spectral representation using windowed Fourier transform using a fast Fourier transform algorithm. Instead of working with simple text we work with the text "spectrograph".

This paper presents a rigorous mathematical proof that if N-grams of a search line are present in the document, then "spectrograms" of both search

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

line and the document are in the same "control" points. A generalized algorithm for full-text search based on the method is proposed, its advantages and disadvantages are analysed.

2. FOURIER WINDOW TRANSFORMATION AND SEARCH BASED ON THE N-GRAMS

Imagine a text as a series of {x (n)} of L length (n = 0,1, ..., L-1). It is advisable to lead it to the lowest register and remove from it all the gaps, punctuation marks and other insignificant symbols [13]. As values of x (n) can be considered for example ASCII-codes of symbols (although you can use your own coding, for example, the number in the range of [-1, 1]). The top graph in Figure 1 shows such a text series of {x (n)}. We divide the text into the N-grams, and as K it is advisable to select numbers as N = 2n, i.e. K = 2, 4, 8.

For simplicity we suppose that the search string consists of one word having a length of N (i.e., the search string consists of one N-gram), which is present in the text and is located beginning from the position n0.

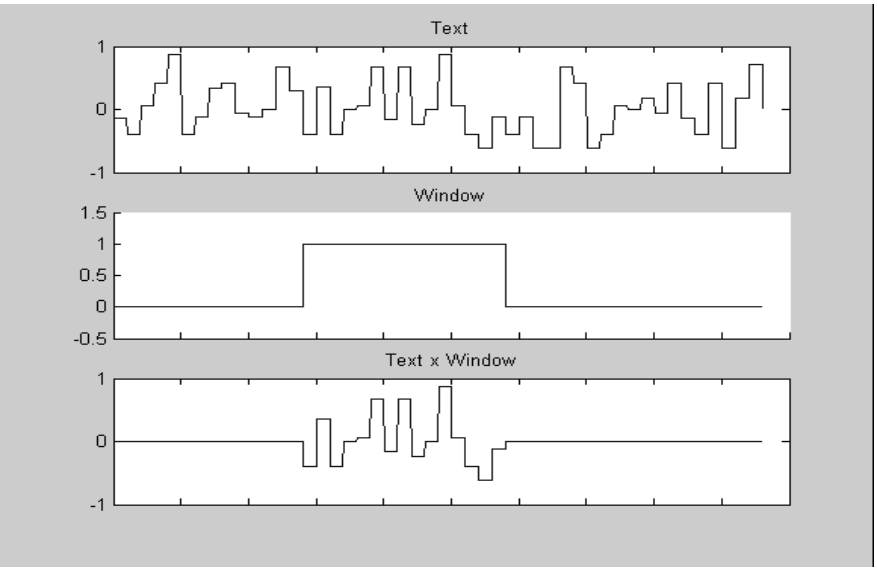

Figure 1: Presentation Of A Text As A Signal {X(N)}

We construct a series for a 'window' function {w (n)} of L length having the zero value everywhere except the interval [n0, n0 +1, ..., n0 + N-1]. This series is shown in the middle picture.

$$w(n)=\begin{cases}1, n\in[n_0,n_0+N-1]\\ 0, n\notin[n_0,n_0+N-1]\end{cases} \quad (1)$$

Product of the series {x (n) · w (n)} (the result of the product is shown in a figure below) allows us to select a keyword (N-gram).

Using Fourier discrete transformation we can compute the spectrum of series {x (n) · w (n)} (n = 0,1, ..., L-1), which can be represented as a de-escalation of the spectra X (p), W (p)

$$\sum_{n=0}^{L-1}x(n)w(n)\ell^{-j\frac{2\pi}{L}kn}= \quad (2)$$
$$=\frac{1}{L}\sum_{p=0}^{L}X(p)W(k-p),k=0,1,,,L-1$$

where X (p) – is a spectrum of the initial text and W (p) – is a spectrum of the 'window' function.

The spectrum of the 'window' function is calculated easily.

$$W(k)=\sum_{n=0}^{L-1}w(n)\ell^{-j\frac{2\pi}{L}kn}=\sum_{n=n_0}^{n_0+N-1}\ell^{-j\frac{2\pi}{L}kn},k=0,1,,,L-1 \quad (3)$$

Making a change of the variable n=p+n0 we write (3) in the form of

$$\sum_{n=n_0}^{n_0+N-1}\ell^{-j\frac{2\pi}{L}kn}=\sum_{p=0}^{N-1}\ell^{-j\frac{2\pi}{L}k(p+n_0)}=\ell^{-j\frac{2\pi}{L}kn_0}\sum_{p=0}^{N-1}\ell^{-j\frac{2\pi}{L}kp} \quad (4)$$

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The expression on the right side of (4) represents the sum of geometric progression which can be found taking into account that a0=1, $q = \ell^{-j\frac{2\pi}{L}k}$:

$$W(k) = \ell^{-j\frac{2\pi}{L}kn_0}\frac{1-\ell^{-j\frac{2\pi}{L}kN}}{1-\ell^{-j\frac{2\pi}{L}k}} =$$

$$\ell^{-j\frac{2\pi}{L}kn_0}\frac{\ell^{-j\frac{\pi}{L}kN}(\ell^{j\frac{\pi}{L}kN}-\ell^{-j\frac{\pi}{L}kN})}{\ell^{-j\frac{\pi}{L}k}(\ell^{j\frac{\pi}{L}k}-\ell^{-j\frac{\pi}{L}k})} = \ell^{-j\frac{2\pi}{L}kn_0}\ell^{-j\frac{\pi}{L}k(N-1)}\frac{\sin\frac{\pi}{L}Nk}{\sin\frac{\pi}{L}k}$$

Thus we obtain

$$W(k) = \ell^{-j\frac{2\pi}{L}kn_0}\ell^{-j\frac{\pi}{L}k(N-1)}\frac{\sin\frac{\pi}{L}Nk}{\sin\frac{\pi}{L}k}, \quad k = 0,1,,,L-1 \tag{5}$$

In turn the left side of the expression (2) can be written as

$$\sum_{n=0}^{L-1}x(n)w(n)\ell^{-j\frac{2\pi}{L}kn} = \sum_{n=n_0}^{n_0+N-1}x(n)\ell^{-j\frac{2\pi}{L}kn} =$$

$$= \sum_{p=0}^{N-1}x(n_0+p)\ell^{-j\frac{2\pi}{L}k(n_0+p)} = \ell^{-j\frac{2\pi}{L}kn_0}\sum_{p=0}^{N-1}x(n_0+p)\ell^{-j\frac{2\pi}{L}kp} \tag{6}$$

We plug in (5) and (6) in (2) and cancel the right and left sides by $\ell^{-j\frac{2\pi}{L}kn_0}$, we obtain basic ratio

$$\sum_{p=0}^{N-1}x(n_0+p)\ell^{-j\frac{2\pi}{L}kp} =$$

$$= \frac{1}{L}\sum_{p=0}^{L}X(p)\ell^{-j\frac{\pi}{L}(N-1)(k-p)}\frac{\sin\frac{\pi}{L}N(k-p)}{\sin\frac{\pi}{L}(k-p)}\ell^{j\frac{2\pi}{L}pn_0} \tag{7}$$

We suppose that in (7) k=0 and thus we have

$$\sum_{p=0}^{N-1}x(n_0+p) =$$

$$= \frac{1}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}\ell^{j\frac{2\pi}{L}pn_0} \tag{8}$$

We introduce a notation

$$Y_0(p) = X(p)\ell^{j\frac{\pi}{L}(N-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p} \tag{9}$$

and write (8) in the following form

$$\sum_{p=0}^{N-1}x(n_0+p) = \frac{1}{L}\sum_{p=0}^{L}Y_0(p)\ell^{j\frac{2\pi}{L}pn_0} = y_0(n_0) \tag{10}$$

The right side of (10) constitutes an expression for Fourier inverse discrete transformation and can be calculated using the inverse Fourier fast conversion algorithm.

We perform similar manipulations for k=L/2 (L must be even, if L is odd, we will add into the text an unlikely and rare symbol, such as '#'), and obtain

$$\sum_{p=0}^{N-1}(-1)^n x(n_0+p) =$$

$$= \frac{\ell^{-j\frac{\pi}{2}(N-1)}}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N-1)p}\frac{\sin(\frac{\pi}{2}N-\frac{\pi}{L}Np)}{\cos\frac{\pi}{L}p}\ell^{j\frac{2\pi}{L}pn_0} = y_{\frac{L}{2}}(n_0) \tag{11}$$

Then we introduce a notation

$$Y_{\frac{L}{2}}(p) = X(p)\ell^{j\frac{\pi}{L}(N-1)p}\frac{\sin(\frac{\pi}{2}N-\frac{\pi}{L}Np)}{\cos\frac{\pi}{L}p} \tag{12}$$

and write (11) in the following form

$$\sum_{p=0}^{N-1}(-1)^n x(n_0+p) =$$

$$= \frac{\ell^{-j\frac{\pi}{2}(N-1)}}{L}\sum_{p=0}^{L}Y_{\frac{L}{2}}(p)\ell^{j\frac{2\pi}{L}pn_0} = y_{\frac{L}{2}}(n_0) \tag{13}$$

which also represents the expression for Fourier inverse discrete transformation and can be calculated using the inverse Fourier fast transformation algorithm.

It can be shown that in calculating of the right sides of expressions (10) and (13), only valid values of complex numbers remain and imaginary components amount to zero. Firstly we need to consider the formula (10) and then write the right side as

$$\frac{1}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}\ell^{j\frac{2\pi}{L}pn_0} =$$

$$= \frac{1}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p} \tag{14}$$

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

We consider a pair of items (p=1,2,…,L/2)

$$X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p} \quad (15)$$

and

$$X(L-p)\ell^{j\frac{\pi}{L}(N+2n_0-1)(L-p)}\frac{\sin\frac{\pi}{L}N(L-p)}{\sin\frac{\pi}{L}(L-p)}=$$

$$=X(L-p)\ell^{j2\pi n_0}\ell^{j\pi(N-1)}\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}N(L-p)}{\sin\frac{\pi}{L}(L-p)} \quad (16)$$

We present $\frac{\sin\frac{\pi}{L}N(L-p)}{\sin\frac{\pi}{L}(L-p)}$ as

$$\frac{\sin\frac{\pi}{L}N(L-p)}{\sin\frac{\pi}{L}(L-p)}=\frac{\sin(\pi N-\frac{\pi}{L}Np)}{\sin(\pi-\frac{\pi}{L}p)}=\frac{\cos\pi N\sin\frac{\pi}{L}Np}{\cos\pi\sin\frac{\pi}{L}p}=(-1)^{N-1}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}$$

and taking into account that $\ell^{j2\pi n_0}=1$, $\ell^{j\pi(N-1)}=(-1)^{N-1}$ и $(-1)^{2(N-1)}=1$ we rewrite the last expression in the following form

$$X(L-p)\ell^{j2\pi n_0}\ell^{j\pi(N-1)}\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}N(L-p)}{\sin\frac{\pi}{L}(L-p)}=$$

$$=X(L-p)\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p} \quad (17)$$

Since X (p) is the spectrum of the actual signal the module is $|X(p)|=|X(L-p)|$ and the phase is $\varphi(p)=-\varphi(L-p)$. We suppose that L is even (if L is odd, we will again add into the text an unlikely and rare symbol, such as '#'), and present (14) as

$$\frac{1}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}=\frac{1}{L}\left\{\begin{array}{l}NX(0)+X(\frac{L}{2})\ell^{j\frac{\pi}{2}(N+2n_0-1)}\sin\frac{\pi}{2}N+\\+\sum_{p=1}^{\frac{L}{2}-1}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}+\sum_{p=1}^{\frac{L}{2}-1}X(L-p)\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}\end{array}\right\}=$$

$$=\frac{1}{L}\left\{\begin{array}{l}NX(0)+X(\frac{L}{2})\ell^{j\frac{\pi}{2}(N+2n_0-1)}\sin\frac{\pi}{2}N+\\+\sum_{p=1}^{\frac{L}{2}-1}|X(p)|\ell^{j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}+\sum_{p=1}^{\frac{L}{2}-1}|X(p)|\ell^{-j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}\end{array}\right\}=$$

$$=\frac{1}{L}\left\{\begin{array}{l}NX(0)+X(\frac{L}{2})\ell^{j\frac{\pi}{2}(N+2n_0-1)}\sin\frac{\pi}{2}N+\\+\sum_{p=1}^{\frac{L}{2}-1}|X(p)|\left(\ell^{j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}+\ell^{-j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}\right)\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}\end{array}\right\}=$$

$$=\frac{1}{L}\left\{\begin{array}{l}NX(0)+X(\frac{L}{2})\ell^{j\frac{\pi}{2}(N+2n_0-1)}\sin\frac{\pi}{2}N+\\+2\sum_{p=1}^{\frac{L}{2}-1}|X(p)|\cos\left(\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right)\frac{\sin\frac{\pi}{L}Np}{\sin\frac{\pi}{L}p}\end{array}\right\}$$

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

First and third items in this expression constitute valid values. X(L/2) is also valid. Therefore it remains to find out that $\ell^{j\frac{\pi}{2}(N+2n_0-1)}\sin\frac{\pi}{2}N$ is a valid value. It is easy to make sure of it finding its imaginary part which (taking into account that $\ell^{j\varphi}=\cos\varphi+j\sin\varphi$) is equal to

$$j\sin\frac{\pi}{2}(N+2n_0-1)\sin\frac{\pi}{2}N=j\cos\pi n_0\sin\frac{\pi}{2}(N-1)\sin\frac{\pi}{2}N=$$
$$=j(-1)^{n_0}\left[\cos\frac{\pi}{2}-\cos\frac{\pi}{2}(2N-1)\right]=0$$

Thus as a result of calculation according to the formula (10), we obtain a valid value.

Now we move to a consideration of formula (13), and write it as

$$\frac{\ell^{-j\frac{\pi}{2}(N-1)}}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin(\frac{\pi}{2}N-\frac{\pi}{L}Np)}{\cos\frac{\pi}{L}p} \qquad (18)$$

In case of even and odd N, different variants are possible. So we want to analyse them separately.

If N is even then

$$\ell^{-j\frac{\pi}{2}(N-1)}=(-j)^{N-1}=(-1)^{N-1}j^{N-1}=-j^{N-1}=$$
$$=j\cdot j^N=j(j^2)^{\frac{N}{2}}=j(-1)^{\frac{N}{2}}=(-1)^{\frac{N}{2}}j$$

$$\sin(\frac{\pi}{2}N-\frac{\pi}{L}Np)=-\cos\frac{\pi}{2}N\sin\frac{\pi}{L}Np=$$
$$=-(-1)^{\frac{N}{2}}\sin\frac{\pi}{L}Np=(-1)^{\frac{N}{2}+1}\sin\frac{\pi}{L}Np$$

If we substitute these expressions into (18) we will get

$$\frac{(-1)^{\frac{N}{2}}j(-1)^{\frac{N}{2}+1}}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}= \qquad (19)$$
$$=-\frac{j}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}$$

As in the analysis of (10) we are going to consider a pair of items, they are

$$X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p} \quad \text{and}$$

$$X(L-p)\ell^{j\frac{\pi}{L}(N+2n_0-1)(L-p)}\frac{\sin\frac{\pi}{L}N(L-p)}{\cos\frac{\pi}{L}(L-p)}=$$

$$=X(L-p)\ell^{j\pi(N-1)}\ell^{j2\pi n_0}\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}N(L-p)}{\cos\frac{\pi}{L}(L-p)}=$$

$$=(-1)^{N-1}X(L-p)\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\left[\pi N-\frac{\pi}{L}Np\right]}{\cos\left[\pi-\frac{\pi}{L}p\right]}=$$

$$=-X(L-p)\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}$$

We use the derived expressions taking into account that $|X(p)|=|X(L-p)|,\varphi(p)=-\varphi(L-p)$ in order to present (18) in the following form:

$$-\frac{j}{L}\sum_{p=0}^{L}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}=-\frac{j}{L}\left\{\begin{array}{l}j\cdot(-1)^{\frac{N}{2}+n_0+1}NX(\frac{L}{2})+\sum_{p=1}^{\frac{L}{2}-1}X(p)\ell^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}\\ -\sum_{p=1}^{\frac{L}{2}-1}X(L-p)\ell^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}\end{array}\right\}=$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$=-\frac{j}{L}\left\{\begin{array}{l} j\cdot(-1)^{\frac{N}{2}+n_0+1}NX(\frac{L}{2})+\sum_{p=1}^{\frac{L}{2}-1}|X(p)|e^{j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p} \\ -\sum_{p=1}^{\frac{L}{2}-1}|X(p)|e^{-j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p} \end{array}\right\}=$$

$$=\frac{j}{L}\left\{j\cdot(-1)^{\frac{N}{2}+n_0+1}NX(\frac{L}{2})+\sum_{p=1}^{\frac{L}{2}-1}X(p)\left[e^{j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}-e^{-j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]}\right]\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}\right\}$$

Taking into account that $\sin\varphi=\frac{e^{j\varphi}-e^{-j\varphi}}{2j}$ we will obtain

$$-\frac{j}{L}\sum_{p=0}^{L}X(p)e^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}=$$

$$=\frac{1}{L}\left\{(-1)^{\frac{N}{2}+n_0}NX(\frac{L}{2})+2\sum_{p=1}^{\frac{L}{2}-1}X(p)\sin\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]\frac{\sin\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}\right\}$$

It is evident that this expression is valid.

If N is odd we will obtain

$$e^{-j\frac{\pi}{2}(N-1)}=(-j)^{N-1}=(-1)^{N-1}j^{N-1}=j^{N-1}=(j^2)^{\frac{N-1}{2}}=(-1)^{\frac{N-1}{2}}$$

$$\sin(\frac{\pi}{2}N-\frac{\pi}{L}Np)=\sin\frac{\pi}{2}N\cos\frac{\pi}{L}Np=(-1)^{\frac{N-1}{2}}\cos\frac{\pi}{L}Np$$

and the expression (18) will be written as

$$\frac{1}{L}\sum_{p=0}^{L}X(p)e^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}$$

In this case a pair of items will look like

$$\frac{1}{L}\sum_{p=0}^{L}X(p)e^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}=\frac{1}{L}\left\{\begin{array}{l} X(0)+X(\frac{L}{2})N+\sum_{p=1}^{\frac{L}{2}-1}X(p)e^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}+ \\ +\sum_{p=1}^{\frac{L}{2}-1}X(L-p)e^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p} \end{array}\right\}=$$

$$X(p)e^{j\frac{\pi}{L}(N+2n_0-1)p}\frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}$$ and

$$X(L-p)e^{j\frac{\pi}{L}(N+2n_0-1)(L-p)}\frac{\cos\frac{\pi}{L}N(L-p)}{\cos\frac{\pi}{L}(L-p)}=$$

$$=X(L-p)e^{j\pi(N+2n_0-1)}e^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}=$$

$$=X(L-p)e^{-j\frac{\pi}{L}(N+2n_0-1)p}\frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p}$$

Taking into account that $|X(p)|=|X(L-p)|,\varphi(p)=-\varphi(L-p)$ we will obtain

Journal of Theoretical and Applied Information Technology
31st July 2014. Vol. 65 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$= \frac{1}{L}\left\{ \begin{array}{l} X(0) + X(\frac{L}{2})N + \sum_{p=1}^{\frac{L}{2}-1} |X(p)| \ell^{j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]} \frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p} + \\ + \sum_{p=1}^{\frac{L}{2}-1} |X(p)| \ell^{-j\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right]} \frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p} \end{array} \right\} =$$

$$= \frac{1}{L}\left\{ X(0) + X(\frac{L}{2})N + 2\sum_{p=1}^{\frac{L}{2}-1} |X(p)| \cos\left[\frac{\pi}{L}(N+2n_0-1)p+\varphi(p)\right] \frac{\cos\frac{\pi}{L}Np}{\cos\frac{\pi}{L}p} \right\}$$

It is evident that this expression is valid. Q.E.D.

We need to find a cross-correlation function between a series {x(n)} with the length L and an N-gram $\{k(n-n_0)\}$ with the length N (n=0,1,…,L-1), where n0 is a replacement of an N-gram relative to the beginning of the text:

$$r(n_0) = \sum_{n=0}^{L-1} x(n)k(n-n_0), \quad n_0 = 0,1,...,L-1$$

Then we need to find Fourier discrete transformation of the series $r(n_0)$. In accordance with the definition

$$R(m) = \sum_{n_0=0}^{L-1} r(n_0)\ell^{-j\frac{2\pi}{L}mn_0} = \sum_{n_0=0}^{L-1}\sum_{n=0}^{L-1} x(n)k(n-n_0)\ell^{-j\frac{2\pi}{L}mn_0} = \sum_{n_0=0}^{L-1}\sum_{n=0}^{L-1} x(n)\ell^{-j\frac{2\pi}{L}mn} k(n-n_0)\ell^{j\frac{2\pi}{L}m(n-n_0)} = X(m)K^*(m)$$

Where K*(m) is a complex coupling to K (m). Here K (m) is the spectrum of N-gram of length N, with a replacement from the beginning of the text.

$$K^*(m) = \sum_{n=0}^{L-1} k(n)\ell^{j\frac{2\pi}{L}mn}$$

Thus

$$R(m) = X(m)K^*(m) \qquad (20)$$

If a text has an N-gram beginning from the position n0 then

$$r(n_0) = \sum_{n=0}^{N-1} k^2(n), \quad n_0 = 0,1,...,L-N-1$$

Therefore if there is a value equal to $\sum_{n=0}^{N-1} k^2(n)$ between the values $\{r(n_0)\}$, the N-gram is present in the text. Values for a given N-gram can be easily found by using the inverse Fourier fast transformation of spectral row $X(m)K^*(m)$ m=0,1,…L-1. This correlation can be a third additional condition of search of the N-grams in the text.

It should be noted that the value $r(n_0)$ at the N-gram will not necessarily be the maximum in the series $\{r(n_0)\}$

3. SPECTRAL SEARCH ENGINE OF CORRESPONDING N-GRAMS

On the basis of the derived correlations (8)-(13) and (20) we can suggest the following algorithm to determine the correspondence of the search string to the lines of a text:

1. Creation of a spectrogram {X(p)} (p=0,1,…L-1, where L is the length of the text) of the initial modified text.

2. Calculation by the formulas (9) and (12) of 'window' and 'N-gram' spectrograms $\{Y_0(m)\}$, $\{Y_{L/2}(m)\}$ and {R(m)} for m=0,1,…L-1, where L is a length of a text and N=4.

$$Y_0(m) = X(m)\ell^{j\frac{\pi}{L}(N-1)m} \frac{\sin\frac{\pi}{L}Nm}{\sin\frac{\pi}{L}m}$$

$$Y_{\frac{L}{2}}(m) = X(m)\ell^{j\frac{\pi}{L}(N-1)m} \frac{\sin(\frac{\pi}{2}N - \frac{\pi}{L}Nm)}{\cos\frac{\pi}{L}m}$$

3. Finding of arrays $\{y_0(n_0)\}$, $\{y_{L/2}(n_0)\}$ n_0 = 0, 1, …, L-N-1 by using the inverse Fourier fast transformation. Carrying out operations in paragraphs 1-3 – these should be undertaken once at the initial stage of processing of a test document. Arrays can be regarded as a spectral analogue of the index of N-grams.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. Formation of a plurality of N-grams from the search string.

5. Chain sampling of N-grams and calculation of the expressions for the next N-gram (length N=4)

$$s = \sum_{n=0}^{N-1} k(n), \quad p = \sum_{n=0}^{N-1} (-1)^n k(n), \quad q = \sum_{n=0}^{N-1} k^2(n), K^*(m) = \sum_{n=0}^{N-1} k(n)\ell^{j\frac{2\pi}{L}mn} \tag{21}$$

where *{k(n)}* are codes of symbols of N-gram of length N.

6. Calculation by the formula (20) of a spectrogram {R(m)} for m=0,1,…L-1 where L is a length of a text and N=4.

$$R(m) = X(m)K^*(m)\ell^{j\frac{\pi}{L}mn_0}$$

7. Finding of arrays $\{r(n_0)\}$ $n_0 = 0, 1, \ldots, L\text{-}N\text{-}1$ by using the inverse Fourier fast transformation.

8. Verification of the conditions

$$\begin{cases} s = y_0(n_0), \\ p = y_{\frac{L}{2}}(n_0), \\ q = r(n_0), \quad n_0 = 0,1,\ldots L - N - 1 \end{cases} \tag{22}$$

If some n0 conditions of (20) are fulfilled then the N-gram is present in the text.

9. Examination of the remaining N-grams. If the correspondence of N-grams from the plurality of inquiry to the strings of the text document does not exceed a certain predetermined threshold quantity, it is considered that the document does not satisfy conditions of the inquiry. Otherwise the document satisfies the inquiry conditions.

In addition the value n0 at which the conditions of (22) are fulfilled determines a place of N-grams in the text that in some cases is important.

4. ASSESSMENT OF EFFICIENCY

As it was said above, this article focuses on the development of new theoretical approach for words searching in the text based on a special, so called "spectral" text representation. Therefore, a detailed comparison of the effectiveness (performance, search speed, error selection) of the proposed method and its "analogs" has not been conducted. It should be noted that currently there are no methods (in any case they are not known to the author) based on a similar ("spectral") approach. Moreover, comparison on artificial test cases will not be representative. Only verification of the proposed algorithm in real search systems, data base management systems will (or will not) show its productivity. Nevertheless, let's obtain some performance measurements.

The effectiveness of the proposed approach can be assessed by comparing the size of an ordinary N-gram index and the sizes of 'spectral' indices. An ordinary N-gram index has a length Ln=L-N+1, and L>>N so we can assume that Ln=L. Each N-gram consists of 4 symbols (for simplicity we assume that for coding of a symbol only 1 byte is used). Consequently keeping custody of the N-gram index requires 4 bytes.

On the other hand, as each of the arrays $\{y_0(n_0)\}$, $\{y_{L/2}(n_0)\}$ has dimension Ls=L-N+1 and L>>N, we also assume that Ls=L. For keeping custody of values $y_0(n_0)$ and $y_{L/2}(n_0)$ we need only one byte, but for the indices we need 2L bytes, i.e. two times less. If keeping custody requires more bytes, memory spans will be commensurable.

It also should be noted that when we use an ordinary N-gram index we need to compare each symbol; when we use the arrays we compare numerical data, which is more efficient in the view of operating speed.

One of the drawbacks of this approach is errors occurring while searching anagrams. They can be called the error of first the kind when coincidence is identified in its absence. For example, the words "dog" and "God" will have the same values of *s, p* and *q* calculated by (21). But in real search fairly long search queries (strings) are requested and the probability of coincidences of a large number of N-grams becomes little. The error of the second kind when the matching is missed is not possible. It should be noted that it is better to give excessive information than to miss important one.

Experimental studies carried out on various sets of documents confirmed the correctness of the above conclusions, efficiency and effectiveness of the proposed algorithm.

5. CONCLUSION

Analysis of modern search engines dealing with texts shows that they are based on text indexing on the basis of N-grams. In itself the procedure of formation of such an index is very cumbersome and requires a lot of resources.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In this paper, we propose a method developed to break the text into N-grams, construct "spectrograms" of the document's text using the windowed Fourier transform and the fast Fourier transform algorithm and then to compare the "spectrogram" of a text document with the "spectrogram" of the search line in special "control" points.

The method works with a spectrogram of the text, which contains the document's content in a qualitatively new form. Semantic notion of N-grams fades into the background. Arithmetic operations are done with spectral values to determine whether the N-gram search line and document text match. Comparison is transferred to the field of comparison of the spectral values on the basis of very simple relations. This method is much easier compared to the symbol-wise comparison.

In order to reduce the search time it is proposed to create "spectral" indices $\{y_0(n_0)\}$, $\{y_{L/2}(n_0)\}$ u $\{r(n_0)\}$ $n_0 = 0,1,\ldots, L\text{-}N\text{-}1$; N=4 for each text document. For a given search inquiry N-grams of length N=4 are built. Consecutively for each N-gram, values are calculated according to the formula (21) and are then compared with the values of the indexes according to the formula (22).

If the conditions of the formula (22) are performed it means that the N-gram is present in the text of the document, and that we can move on the verification of the following N-grams. If the amount of correspondences of N-grams from the plurality of inquiry to the strings of the text document does not exceed a certain predetermined threshold quantity, it is considered that the document does not satisfy the inquiry conditions. Otherwise the document satisfies the inquiry conditions and the document is included in the resulting list of the search engine.

In some cases there may be errors of the first kind. But in the real search long search strings are usually given and the likelihood of coincidences of a huge amount of N-grams becomes small. Errors of the second kind are impossible.

All the derived correlations have a pure mathematical proof and passed a pilot test in the system MatLab 6.5, which confirmed their full correctness.

In the future, it is advisable to verify the effectiveness of the proposed algorithm in real search engines and databases (e.g. in PostgreSQL). We hope such tests will be performed by those who are interested in the proposed method.

REFRENCES:

[1] Witten, I.H., Moffat, A., Bell, T.C. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*. 2nd ed. San Francisco: Morgan Kaufmann.

[2] Kim M.S. et al. 2008. 'Structural optimization of a full-text n-gram index using relational normalization'. *The VLDB Journal*. 17(6), pp.1485-1507.

[3] Zobel, J., Moffat, A.. 2006. 'Inverted files for text search engines'. *ACM Comput Surv*. 38(2), Article No. 6.

[4] Mäkinen V. 2003. 'Compact suffix array–A space-efficient full-text index'. *Fundamenta Informaticae*. 56(1), pp. 191-210.

[5] Farach, M. 1997. Optimal suffix tree construction with large alphabets. In: *Proceedings of 38th Annual Symposium on Foundations of Computer Science*. Washington DC: IEEE, pp.137-143.

[6] Baeza-Yates, R., Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. New York: ACM Press.

[7] Mayfield, J., McNamee, P. 2003. Single N-gram stemming. In: *Proceedings of 26th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval*, pp. 415-416.

[8] Baeza-Yates, R., Navarro, G. 1998. A practical q-gram index for text retrieval allowing errors. CLEI [Online] 1(2). Available from: http://callisto.nsu.ru/documentation/CSIR/selected/ngramm/i2p3/i2p3.pdf

[9] Miller, E., Shen, D., Liu, J., Nicholas, C. 2000. 'Performance and scalability of a large-scale N-gram based information retrieval system'. *Journal of Digital Information* [Online]. 1(5), pp.1–25. Available from: http://users.soe.ucsc.edu/~elm/Papers/jodi00.pdf

[10] Miller, E., Shen, D., Liu, J., Nicholas, C. 'Performance and scalability of a large-scale N-gram based information retrieval system'. *Journal of Digital Information*. 1(5), pp.1-25.

[11] Scholer, F., Williams, H.E., Yiannis, J., Zobel, J. 2002. Compression of inverted indexes for fast query evaluation. In: *Proceedings of International Conference on Information Retrieval*, SIGIR, ACM pp. 222-229.

[12] Sutinen, E., Tarhio, J. 1996. Filtration with q-samples in approximate string matching. In: *Proceedings of 7th Annual Symposium on*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Combinatorial Pattern Matching (CPM). Lecture Notes in Computer Science 1075, Springer, Berlin, pp. 50-63.

[13] Kudriavtsev K.Y., Korotkov A.E. (2012). *Methods for increasing the speed of information retrieval in databases*. Lambert Academic Publishing.

[14] Sergienko A.B. 2006. *Digital signal processing*. St-Petersburg: Piter.