# A Text Scanning Mechanism Simulating Human Reading Process

**Bei Xu, Hai Zhuge***
Nanjing University of Posts and Telecommunications, China
Key Lab of Intelligent Information processing, Chinese Academy of Sciences, China
xubei@njupt.edu.cn, zhuge@ict.ac.cn

## Abstract

Previous text processing techniques focus on text itself while neglecting human reading process. Therefore they are limited in special applications. This paper proposes a text scanning mechanism for generating the dynamic impressions of words in text by simulating recall, association and forget processes during reading. Experiments show that the mechanism is suitable for multiple text processing applications.

## 1 Introduction

Previous text processing techniques focus on retrieve, extract, summarize, mine, analyze and organize symbols in text. They can be classified into three kinds. The first is based on the features of text [Radev *et al*., 2002], including term frequency [Luhn, 1958], positions of words or sentences [Baxendale, 1958], etc. *TF-IDF* is a typical technique based on term frequency [Salton and Buckley, 1988]. Some applications are implemented by using *TF-IDF* [Wu *et al*., 2008]. Some techniques represent text as a network and apply network methods to process text [Costa *et al*., 2007]. The second is based on training. Various models or schemas are designed and trained by well-defined data. The typical techniques include vector space models [Singhal, 2001], latent semantic analysis [Dumais, 2004], probabilistic model [Amati and Rijsbergen, 2002; Lafferty and Zhai, 2003]. The training process was improved by using natural language instead of labeled examples [Goldwasser and Roth, 2011]. The third incorporates pre-established frames or background knowledge, such as wordnet [Fellbaum, 2010], conceptual knowledge [Musat *et al*., 2011], ontology knowledge [Batet *et al*., 2011], encyclopedic knowledge [Nastase, 2008], etc.

However, previous techniques are only suitable for specific applications. A key shortcoming of previous research methods is that they neglect human reading process. Newell pointed out that the natural processes, including cognition process, can be simulated and the information in the simulations can deal with any issues about the processes [Newell, 1990].

The motivation of this paper is to build a text processing mechanism suitable for multiple applications by simulating human reading process.

From a broader perspective, human cognitive aspect has been considered as a general mechanism to effectively process information in multiple areas. A model of information foraging based on the constraints and mechanisms of human cognitive architecture was developed [Pirolli and Fu, 2003]. The decay rate of cognition was incorporated for information storage in a cognitive architecture [Anderson, 2007]. A cognitive network was used to solve complex problem based on the simulation of decomposing complex task [Bhattacharyya and Ohlsson, 2010]. The spread of information through a network of individuals based on the simulation of interaction between individual cognitive mechanisms and social dynamics was studied [Coman *et al*., 2012]. The agents with human-like memory which can forget and maintain information were designed [Reitter and Lebiere, 2012]. The advantages of solving problems by collecting information from human actions were reported [Kapoor, 2012].

Zhuge pointed out that reading is a process of constructing semantic link networks of concepts in reader's mind by discovering and browsing the semantic link networks of words weaved by the author [Zhuge, 2012]. A Semantic Link Network model was developed as a general method and theory for self-organized semantic networking. It was extended to semantically link objects in different spaces to study the fundamental structure of cyber-physical society and create cyber-physical-social intelligence [Zhuge, 2011]. A semantic lens mechanism based on the semantic link network and a multi-dimensional classification space was proposed to simulate semantic networking mechanism and multi-dimensional information processing mechanism of human cognition and behavior [Zhuge, 2010]. Human reading processes were considered for implementing faceted navigation on text [Xu and Zhuge, 2012a; Xu and Zhuge, 2012b].

Cognition scientists proposed some models to explain human reading process [Fauconnier, 2002], but they have not proposed any applicable method.

This paper proposes a text scanning mechanism simulating human reading process (*HTSM*). It inputs a text and then outputs dynamic impressions of words generated through scanning the text. It has the following characteristics:

1. Generality. The dynamic impressions of words can be used in different applications as stated in section 3.
2. Dynamicity. *HTSM* scans and analyzes text sentence by sentence from beginning to end to generate the dynamic impressions of the words in the text.
3. Without training or background knowledge request.

## 2 The Text Scanning Mechanism

### 2.1 Global and Local Word Impressions

Assuming that the scanning text has $N$ sentences, and the scanning sentence is the $k^{th}$ sentence within a local range of size $D$ ($0<D\leqslant N$). The global range refers to the range from the first sentence to the $k^{th}$ sentence. The local range refers to the range from

1. the first sentence to the $(k+D)^{th}$ sentence if ($k\in[1, D]$),
2. the $(k–D)^{th}$ sentence to the $(k+D)^{th}$ sentence if ($k \in [D+1, N–D]$),
3. the $(k–D)^{th}$ sentence to the *Nth* sentence if ($k \in [N–D+1, N]$).

The left part of Figure 1 is an example of the global range and local range.

Readers have impressions of words, which keep changing while reading. Two kinds of impressions of words are distinguished in *HTSM*: Global Word Impression (*GWI*) and Local Word Impression (*LWI*). $GWI_k(i)$ represents the impression of word $i$ generated within the global range while scanning the $k^{th}$ sentence. $LWI_k(i)$ represents the impression of word $i$ generated within the local range while scanning the $k^{th}$ sentence. *GWI* and *LWI* are dynamic because *HTSM* constantly gets *LWI*s of words from the moving local range and accumulate the *LWI*s on the *GWI* of the words during scanning. The initial *GWI*s of words are set as 0.

*HTSM* calculates and records *GWI*s and *LWI*s of all words during scanning.

### 2.2 Words' Global Relevancy and Local Relevancy

The calculation of *GWI*s and *LWI*s is based on the relevancy between words. Empiricists like David Hume think that people believe things are relevant when the things frequently appear in succession at logic, space, or time. In line with this idea, two words are regarded as relevant within a text if the two words frequently emerge in a common text unit (e.g., phrase, sentence, paragraph, chapter). We regard sentences as the basic units because words in a common sentence have greater relevancy than they are in different sentences. The following are two kinds of relevancy between words.

**Definition 1.** The global relevancy between two words, $w_1$ and $w_2$, $GR_k(w_1, w_2)$, represents the relevancy between the two words in the global range while scanning the $k^{th}$ sentence. It is measured by the number of sentences containing $w_1$ and $w_2$ in the global range.

Human also focus on local range while reading. So, two words may be weakly relevant in a local range while they have strong global relevancy. For example, given a text on play "*Romeo and Juliet*" and two words "*Romeo*" and "*Juliet*" have strong global relevancy, if there is a local range that only introduces Romeo and his families, the relevancy between "*Romeo*" and "*Juliet*" is weak within the local range. The local relevancy is defined as follows:

**Definition 2.** The local relevancy between two words $w_1$ and $w_2$, $LR_{k,D}(w_1, w_2)$, represents the relevancy between the two words in the local range while scanning the $k^{th}$ sentence. $LR_{k,D}(w_1, w_2)$ is calculated by the following equations:

$$LR_{k,D}(w_1, w_2) = Min(N_{k,D}, 1) \times GR_k(w_1, w_2) \quad (1),$$

$$N_{k,D} = \begin{cases} \dfrac{NumS_{k-D,k+D}(w_1, w_2)}{\log_r(NumS_{1,k}(w_1, w_2))} & \text{if}(k \in [D+1, N-D]) \\ 1 & \text{if}(k \in [1, D]) \\ \dfrac{NumS_{k-D,N}(w_1, w_2)}{\log_r(NumS_{1,N}(w_1, w_2))} & \text{if}(k \in [N-D+1, N]) \end{cases} \quad (2),$$

where $N_{k,D}$ denotes the ratio of local relevancy and global relevancy; $NumS_{x,y}(w_1,w_2)$ denodes the number of sentences containing $w_1$ and $w_2$ from the $x^{th}$ sentence to the $y^{th}$ sentence; *Min*() is a function that returns the smaller value; and $r$ is an integer.

Equaltion 2 represents that the more times two words emerge within a local range, the bigger *LR* the two words have. The maximum value of two words' local relevancy is limited by the corresponding global relevancy. Meanwhile, the increase of local relevancy is fast if the corresponding global relevancy is high.

### 2.3 Two-Layer Word Network

*HTSM* is performed on a two-layer word network.

One layer is Global Word Network, formally described as $GWN_k(V_k, E_k, GWI_k[], GR_k[])$, where $k$ denotes the scanning sentence. $V_k$ denotes a set of words from the global range. Two words $w_1$ and $w_2$ have an undirected link if $GR_k(w_1, w_2)$ is bigger than zero and $E_k$ denotes the set of links. Every link in $GWN_k$ has a weight that equals to the corresponding global relevancy and $GR_k[]$ records the links' weights. Every word in $GWN_k$ has a weight within $[0,+\infty)$ that equals its *GWI* and $GWI_k[]$ records the words' weights.

The other layer is Local Word Network, formally described as $LWN_k$ ($V_k$, $E_k$, $LWI_k[]$, $LR_k[]$, $D$). $k$ and $D$ denote the local range. $V_k$ in $LWN_k$ denotes the set of words in the local range. There is an undirected link between two words $w_1$ and $w_2$ if $LR_{k,D}(w_1, w_2)$ is bigger than zero and $E_k$ in $LWN_k$ denotes the set of links. Every link in $LWN_k$ has a weight that equals to the corresponding local relevancy and $LR_k[]$ records the links' weights. Every word in $LWN_k$ has a weight within $[0,+\infty)$ that equals its *LWI* and $LWI_k[]$ records the words' weights.

*GWN* and *LWN* are fully connected. Figure 1 shows an example of a two-layer word network while scanning the $k^{th}$ sentence.

In *HTSM*, only nouns are considered because nouns directly reflect objects while the other types of words render objects.

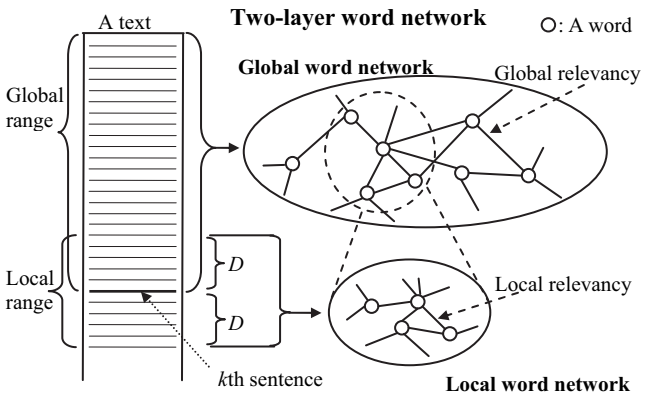


Figure 1. A two-layer word network generated from global range and local range.

## 2.4 Recall, Association and Forget Processes

Three processes occur while scanning: recall process, association process and forget process.

The previous memory of a word will be recalled if it is mentioned. So, a word's previous *GWI* will be regained in the recall process if a new sentence contains the word. Given a sentence which is the $k^{th}$ sentence and contains a word $w$, the *GWI* of $w$ will reach the maximum value of *GWI*s in previous scanning process. The steps of recall process are as follows, where *Max*() is a function that returns the biggest value.

**The recall process of $k^{th}$ sentence.**
**Input:** $GWI_{k-1}[]$—*GWI*s of nodes in $GWN_{k-1}$, $s$—the $k^{th}$ sentence
**Output**: $RP_{k-1}[]$—*GWI*s of nodes in $GWN_{k-1}$ after recall process
**Steps:**
For each element $GWI_{k-1}(i)$ in $GWI_{k-1}[]$
  If (word $i$ is in $s$)
    $RP_{k-1}(i)= Max(GWI_1(i), GWI_2(i),\ldots, GWI_{k-1}(i))$;
  Else
    $RP_{k-1}(i)= GWI_{k-1}(i)$;
Return $RP_{k-1}[]$

The maximum *GWI* will be regained after several times of appearances of the word. The recall process of words within a text is very fast during reading. So, the appearance times of regaining a word's maximum *GWI* is set as 1.

In the association process, words' *LWI*s influence each other through words' local relevancy. If two words have local relevancy and one of them is mentioned in text, human mind will associate the other word and the *LWI*s of both words are enhanced. For example, given a text describing Romeo and Juliet, the two words "*Romeo*" and "*Juliet*" have local relevancy, so that the *LWI* of "*Juliet*" increases once "*Romeo*" is mentioned. Figure 2 shows the association process of the $k^{th}$ sentence by using the text on the second world war (http://en.wikipedia.org/wiki/WWII).

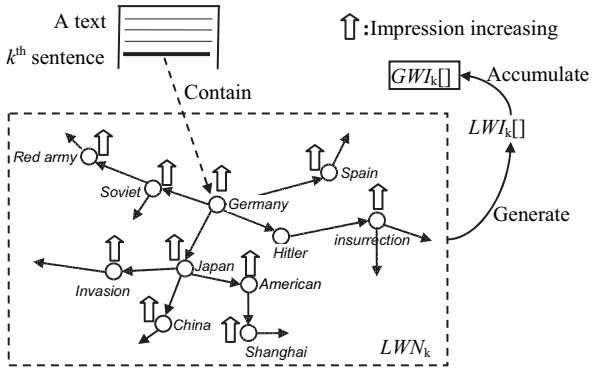


Figure 2. The association process caused by the word "*Germany*" in the $k^{th}$ sentence. The weights are propagated through the arrows. $LWI_k[]$ is generated from the weight propagation and is accumulated on $GWI_k[]$.

Assuming that the $k^{th}$ sentence which contains words $w_1$, $w_2$,…, and $w_n$, the association process is described as follows, where $\beta$ is set as the number of nouns in the $k^{th}$ sentence.

**The association process of the $k^{th}$ sentence.**
**Input:** the $k^{th}$ sentence containing words $w_1,w_2,\ldots,w_n$
**Output:** $LWI_k[]$
($W[i]$ denotes the weight of word $i$)
**Steps:**
Set nodes in $LWN_k$ unassociated and their weights as 0.
Set two node sets $T$ and $R$;
For each nodes $t$ in $\{w_1,w_2,\ldots,w_n\}$
  $W[t] = W[t] +\beta$;
  Add $t$ into $T$ and set $t$ associated;
While ($T$ is not empty)
  For each node $i$ in $T$  //find the nodes that will receive weight
    If (the weight of $i$ is bigger than *MIN*)
      For each node $j$ in the neighbors of $i$
        If ($j$ is unassociated)
          Add $j$ into $R$;
          Set $j$ associated;
  For each node $i$ in $T$  //propagate weight
    For each node $j$ in $R$
      If ($LR_k(i,j)>0$)
        $W[j] = W[j]+ PW(i\rightarrow j)$;
        //$PW(i\rightarrow j)$ is calculated in equation 3.
      Else continue;
    $W[i] = W[i]*(1/\omega)$; // $\omega$ is remain ratio.
  $T=R$;
  Clear $R$;
Output LWIk[] as the weights of nodes W[];

$$PW(i\rightarrow j)=\begin{cases}(1-1/\omega)\times W[i]\times\dfrac{LR_{k,D}(i,j)}{\sum_j LR_{k,D}(i,j)} & if\ (W[i]>MIN)\\ 0 \quad if\ (W[i]\le MIN)\end{cases} \tag{3}$$

In equation 3, $j$ denotes a neighbor word of $i$; $PW(i\rightarrow j)$ denotes the weight propagated from $i$ to $j$; $LR_{k,D}(i,j)$ denotes the local relevancy between $i$ and $j$ after scanning the $k^{th}$ sentence. *MIN* is a threshold that the propagated weight lower than *MIN* will be ignored. $\omega$ is a remain ratio that a word will keep $1/\omega$ weight and propagate $(1-1/\omega)$ weight to neighbors. $\omega$ is usually set as 2.

Forget process refers to the following phenomenon: A word's *GWI* gradually fades out if the word is not mentioned in a certain time.

The following are three factors determining a word's forget process: (1) *Forget time*, which refers to the number of sentences from the sentence containing the word scanned last time to the scanning sentence. For example, assuming that the $i^{th}$ sentence and the $(i+j)^{th}$ sentence contain a common word $w$ and the sentences between the two sentences do not contain $w$, then $w$'s forget time is $j-1$ when *HTSM* scans the $(i+j-1)^{th}$ sentence and is 0 when *HTSM* scans the $(i+j)^{th}$ sentence. The bigger the forget time is, the smaller *GWI* of a word has. (2) *Repeating times*, which refers to the number of sentences containing the word before the scanning sentence. Words will not be forgotten quickly if it has been mentioned many times. (3) *Link diversity*, which refers to the number of links between the word and its neighbors. In psychology, a thing is hardly forgotten if it is linked to many other things and easy to fade out if it has few links to other things. Therefore, in *GSTM*, a word will be hardly forgotten if its link diversity is high and will quickly fade out if its link diversity is low.

Combined the three factors, the forget process of word $i$ is calculated by the following equations:

$$GWI_k(i) = C_{Forget} \times GWI_{k-1}(i) = (1 - \frac{1}{\lambda \times (1 + (FT - \alpha)^2)}) \times GWI_{k-1}(i) \qquad (FT \geq 1) \tag{4},$$

and

$$\alpha = Max(MIN_\alpha,\ \log_2 \frac{\prod_j (1 + GR_{k-1}(i,j))}{NL_k(i) + \sum_j GR_{k-1}(i,j)}) \tag{5}.$$

In equation 4, $FT$ denotes the forget time; $C_{\text{Forget}}$ is the coefficient that reflects the speed of forget process. $\lambda$ determines the lowest value of $C_{\text{Forget}}$. For example, $C_{\text{Forget}}$ is within [1/2, 1) when $\lambda$=2 and $C_{\text{Forget}}$ is within [1/4, 1) when $\lambda$=4. If $FT$=0, the forget process does not occur. In equation 5, $NL_k(i)$ denotes the number of links between $i$ and its neighbors after scanning the $k^{th}$ sentence. There is a smallest value of $\alpha$ (denoted as $MIN_\alpha$) because people do not forget things immediately after seeing them. The usual value of $MIN_\alpha$ is 5.

Equation 4 depicts that the impression of a word fades slowly at first, then become faster with the growth of $FT$, and slows down at last. Figure 3 shows the change of $C_{\text{Forget}}$ with different $\alpha$. The bigger $\alpha$ is, the bigger $FT$ is needed to get into fast forget speed. So the impression of a word fades slowly with the growth of $\alpha$.

Equation 5 implies the influence of *repeating times* and *link diversity* because of the following reasons:

1. Once a sentence contains word $i$, the global relevancies between $i$ and some of its neighbors will increase. In equation 5, the numerator increases much faster than the denominator when some links' global relevancies are enhanced. Therefore $\alpha$ increases when *repeating times* increases.
2. Given a fixed value of $\Sigma GR_k(i, j)$ in equation 5, the more links word $i$ has, the bigger $\Pi(1+GR_k(i, j))$ is and it grows much faster than $NL_k(i)$. Therefore $\alpha$ increases when *link diversity* increases.

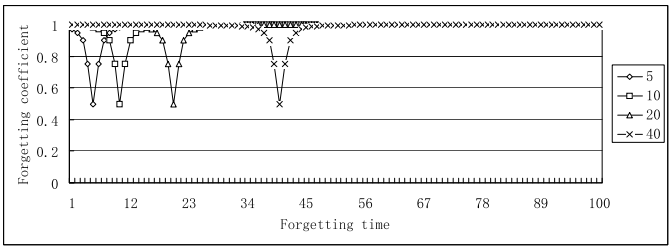


Figure 3. The curve of $C_{\text{Forget}}$ when $\alpha$=5, $\alpha$=10, $\alpha$=20 and $\alpha$=40. $r$=1.5, $\lambda$=2.

Combined the three processes in *HTSM*, the integrated calculation is shown in Figure 4.

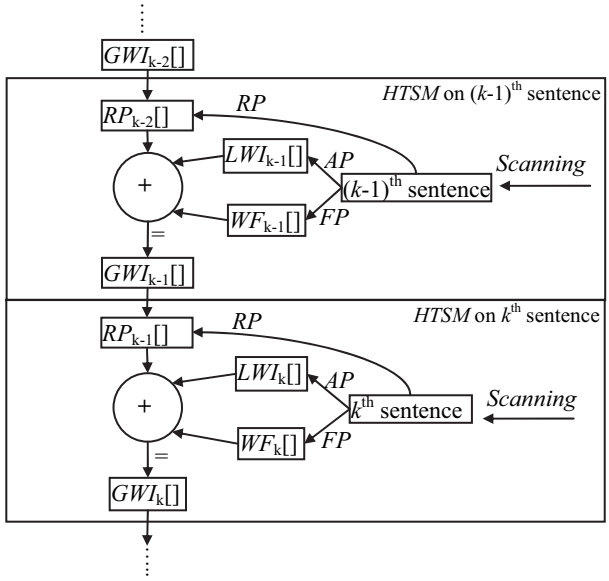


Figure 4. The co-influence of recall process, association process and forget process. *AP* denotes association process, *FP* denotes forget process and *RP* denotes recall process. $RP_{k-1}[]$ and $RP_k[]$ denote the *GWI*s after recall process. $LWI_{k-1}[]$ and $LWI_k[]$ denote the variations of weights of words from association process. $WF_{k-1}[]$ and $WF_k[]$ denote the variations of weights of words from the forget process. $GWI_k[]$ denotes the weights of words in $GWN_k$; $GWI_{k-1}[]$ denotes the weights of words in $GWN_{k-1}$; $k \in [1,N]$ and $N$ is the total number of sentences.

Equation 6 calculates words' *GWI*s. Notice that the values in $LWI_k[]$ are greater than or equal to 0 and the values in $WF_k[]$ are less than or equal to 0.

$$GWI_k[] = \begin{cases} RP_{k-1}[] + LWI_k[] + WF_k[] & (k>1) \\ LWI_k[] & (k=1) \end{cases} \tag{6},$$

### 2.5 *GWI* Curve and *GWIV* Curve

A word's global word impression curve (*GWIC*) is a curve that records the word's *GWI*s under the co-influence of recall process, association process and forget process. Figure 5 shows an example of $GWIC_{\text{hitler}}$ on word "*Hitler*" from an article in http://en.wikipedia.org/wiki/WWII. The remaining *GWI* of a word after a certain number of sentences without mentioning the word increases with the mention times of the word. For example, in Figure 5, the *GWI*s around *B* are bigger than the *GWI*s around *A*. Based on *GWIC*, global word impression variation curve (*GWIVC*) depicts the variation of *GWIC* and can be defined as follows:

$$GWIVC_i(k)=\begin{cases} LWI_k(i) & \text{(if word } i \text{ is in } k^{th} \text{ sentence)} \\ GWIC_i(k)-GWIC_i(k\text{-}1) & \text{(if word } i \text{ is not in } k^{th} \text{ sentence)} \end{cases} \quad (7),$$

Where $GWIVC_i(k)$ denotes the value of *GWIVC* on word *i* after scanning the $k^{th}$ sentence.

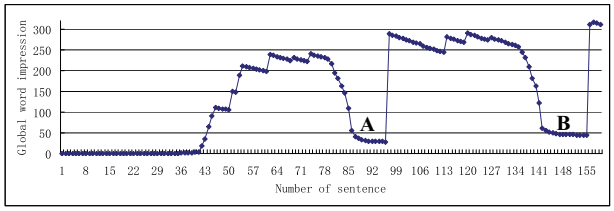


Figure 5. $GWIC_{\text{Hitler}}$ on the first 160 sentences. $r$=1.5, $D$=20, $\omega$=2, $\lambda$=2.

Figure 6 illustrates an example of *GWIVC* on word "*Hitler*" in the article.

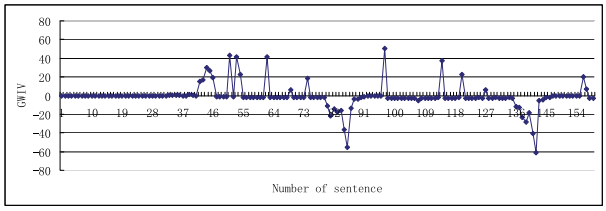


Figure 6. $GWIVC_{\text{Hitler}}$ on the first 160 sentences.

The calculation of *HTSM* is fast because: (1) it is not iterative and each weight propagation process will terminate soon; and, (2) it does not need re-scanning the previous sentences while scanning a new sentence. The time complexities of recall process and forget process are less than the time complexity of association process. So, the upper bound of time complexity of *HTSM* is $O(3*N*W*\log_{\omega}(W/MIN))$, where *N* is the total number of sentences in a text; *W* is the number of words in the longest sentence in the text; $\omega$ and *MIN* are the coefficients in equation 3. *W*, $\omega$ and *MIN* can be approximately considered as constant. So, the time complexity is approximately linear and is acceptable in most applications.

## 3 Applications

### 3.1 Information Extraction Based on Keywords

A useful application of *HTSM* is information extraction. Users can obtain needed information from text by inputting several keywords representing their preferences. This issue can be handled by the comparison of different contributions from each sentence to the keywords' *GWIVC*. A method based on *HTSM* is as follows:

1. Calculate the *GWIVC* of every keyword.
2. Obtain the integrated *GWIVC* on the keyword set. The horizontal coordinates of integrated *GWIVC* are sentences; the longitudinal coordinates are summation of the values in every keywords' *GWIVC*s.
3. Find the first *N* values in the integrated *GWIVC*s and extract the corresponding sentences as output (*N* is the extraction ratio that can be adjusted by users).

An experiment is designed to verify the effectiveness of the method. 100 pieces of news on different topics are randomly chosen from *DUC*05 and *DUC*06 (http://www-nlpir.nist.gov/projects/duc/data.html). For each piece of news, five nouns with the top-five highest frequencies are chosen to be the keywords of the piece of news. Each piece of news can be considered as an ideal result of an extraction by using its corresponding keywords. 100 extractions are performed. The extraction ratio of an extraction is set as *n* while extracting a piece of news containing *n* sentences. The match ratio of an extraction is defined as: *MR = the number of common sentences in the ideal result and the extracted result* / *the number of sentences in the ideal result*. Figure 7 shows the *MR*s of 100 extractions.

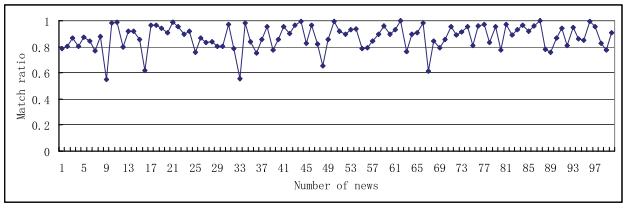


Figure 7. The match ratios. $r$=1.5, $D$=20, $\omega$=2, $\lambda$=2.

In Figure 7, most values are above 80% and the average value is 87.3%. The extracted pieces of news basically match the ideal results. So, the method of information extraction based on *HTSM* is effective.

In *HTSM*, different values of *D* lead to different effects. The effect can be studied by setting different values. Considering the following cases:

(1) *D* equals to the number of sentences in a text. Words' weights are inclined to be propogated between words with high global relevancies. It may be inappropriate with the meaning of text. For example, given a text that contains 200 sentences and describes three person *A*, *B* and *C*, the first 90 sentences describe the things happened between *A* and *B*, 91-120 sentences describe things happened between *A* and *C*, 121-200 sentences describe things happened between *A* and *B*. So, when a sentence mentions *A* within 91-120 sentences,

the weight of *A* should mainly flow to *C* compared to *B*. However, if *D* equals to the number of sentences, the weight from *A* will mainly flow to *B* according to equation 3. More generally, the inappropriate weight flow occurs when *D* is too big to reflect the content's meaning.

(2) *D* equals to 1. In this case, weights is only propagated between words in common sentences which limits the association process. The inappropriate weight flow occurs when *D* is too small.

A suitable *D* exists between the two extreme situations. We perform an experiment on the first 20 pieces of news with different values of *D*. Figure 8 shows the average match ratios while giving *D* a value from 1 to 400. We can see that the average match ratios are higher than others when *D* approximately equals 20. The best result is 92.3% when *D* equals to 18. The average match ratios become low, unstable and irregular when *D* is from 50 to 400 because *D* may or may not reflect the meaning of text. Figure 9 shows the results of another experiment on re-picked 20 pieces of news. The best result is 91.7% when *D* equals to 25. Therefore, different values of *D* are suitable for different text and generates different effects for an application. *D* can be adjusted to reach the best result in an application.

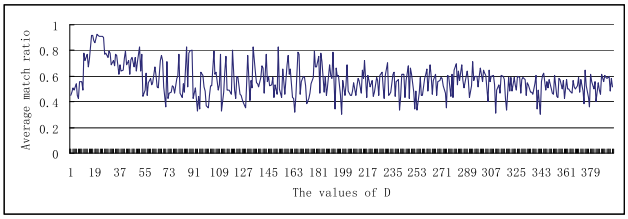


Figure 8. The average match ratios while setting different values to *D* on the first 20 pieces of news.

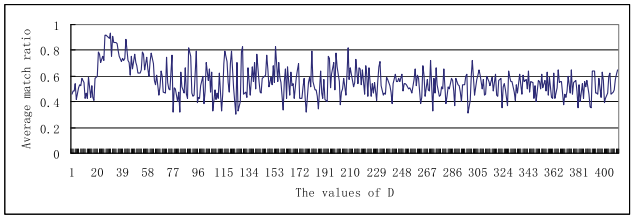


Figure 9. The average match ratios while setting different values to *D* on re-picked 20 pieces of news.

## 3.2 Keyword Extraction

Keyword extraction is to extract important words from text. In *HTSM*, the importance of a word is measured by the summation of its *GWI*s as follows:

$$IM_k(i) = \sum_{j=1}^{k} GWI_j(i) \qquad (8),$$

Where $IM_k(i)$ denotes the importance of word *i* after scanning the $k^{th}$ sentence. Thus, the keywords can be extracted by sorting their importance.

The proposed method has the following characteristics: (1) The extracted keywords change with the scanning process. (2) The importance of a word increases when the frequency of the word increases. (3) Given a text and two words *i* and *j* with same frequency, despite other factors, *i* should be more important than *j* if *j* is concentrated in some areas within the text and *i* is evenly distributed over the text. Equation 8 implies the characteristic because the recall process on *i* will bring more importance compared to the recall process on *j*. (4) The importance of a word increases when the local relevancy between the word and its neighbors increase because the word receives weight from other words through association process.

An experiment is designed to demonstrate the effectiveness of the proposed method. Three groups of text are chosen. Group 1 includes 20 technical papers on AI; group 2 includes 20 web pages from Wikipedia on different topics; group 3 includes 20 pieces of news from *DUC*06 on different topics. Annotators manually label each article with 15 keywords which is considered as the ideal results. The label words must be nouns in the text. Three methods are tested: *tf*, *tf-idf*, and ours. *tf* denotes the method that extracts keywords according to their term frequencies. *tf-idf* can distinguish a word's importance to a text in a set of texts. In the experiment, *tf-idf* is separately performed on the three groups. 15 keywords are extracted for each text. The match ratio of keywords of a text is defined as: $MR_{\text{keyword}}$ = *the number of words in the intersection of the ideal result and the extracted keywords* / *the number of words in the ideal result*. Figure 10 shows the results of the experiment.

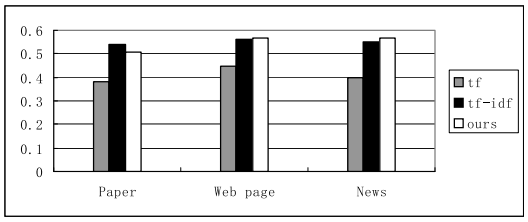


Figure 10. The average $MR_{\text{keyword}}$ of three groups by using different methods.

The experiments show that *HTSM* is much better than *tf* and is similar to *tf-idf*. So it is effective for keyword extraction. Compared to *tf-idf*, *HTSM* can eliminate keywords with concentrated distributions because of the third characteristic and discover important keywords with low frequencies because of the fourth characteristic.

## 4 Conclusions and Future Work

This paper proposes a text scanning mechanism for generating the dynamic impressions of words in text by simulating human reading process. It incorporates recall, association and forget processes involved in reading. Experiments show that it is effective in different applications. The mechanism provides a new method for intelligent text processing. We are working on a methodology for applying it to general text processing applications.

## Acknowledgement

Research supported by National Science Foundation of China (61075074) and fundings from Nanjing University of Posts and Telecommunications. * Correspondence author.

## References


[Amati and Rijsbergen, 2002] G. Amati, C. J. V. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4), 357–389, 2002.

[Anderson, 2007] J. R. Anderson. *How can the human mind occur in the physical universe?* Oxford University Press, Oxford, UK, 2007.

[Batet *et al*., 2011] M. Batet, D. Sanchez, A. Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*. 44(1), 118–125, 2011.

[Baxendale, 1958] P. B. Baxendale. Man-made index for technical literature—an experiment, *IBM Journal of Research and Development*, 2(4), 354-361, 1958.

[Bhattacharyya and Ohlsson, 2010] S. Bhattacharyya and S. Ohlsson. Social creativity as a function of agent cognition and network properties: A computer model. *Social Networks*, 32(4):263–278, 2010.

[Coman et al., 2012] A. Coman, A. Kolling, M. Lewis and W. Hirst. Mnemonic convergence: from empirical data to large-scale dynamics, In *Proceedings of the International Conference on Social Computing*, *Behavioral- Cultural Modeling and Prediction*, 7227, 256-265, 2012.

[Costa *et al*., 2007] L. D. F. Costa, F. A. Rodrigues, G. Travieso and P. R. V. Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1), 167-242, 2007.

[Dumais, 2004] S. T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38, 189-230, 2004.

[Fauconnier, 2002] G. Fauconnier and M. Turner. *The Way We Think*: *Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books, 2002.

[Fellbaum, 2010] Christiane Fellbaum. WordNet. *Theory and Applications of Ontology*: *Computer Applications*, 231-243, 2010.

[Goldwasser and Roth, 2011] D. Goldwasser and D. Roth. Learning from Natural Instructions. In *Proceedings of the 22th International Joint Conference on Artificial Intelligence*, 1794-1800, 2011.

[Kapoor, 2012] A. Kapoor, B. Lee, D. Tan and E. Horvitz. Learning to Learn: Algorithmic Inspirations from Human Problem Solving. In *Proceedings of 26th AAAI Conference on Artificial Intelligence*, 1571-1577, 2012.

[Lafferty and Zhai, 2003] J. Lafferty, C. Zhai. Probabilistic relevance models based on document and query generation. *Language Modeling and Information Retrieval*. W. B. Croft and J. Lafferty Eds., Kluwer Academic Publishers, 2003.

[Luhn, 1958] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research Development*. 2(2): 159-165, 1958.

[Musat *et al*., 2011] C. C. Musat, J. Velcin, S. Trausan-Matu and Marian-Andrei Rizoiu. Improving Topic Evaluation Using Conceptual Knowledge. In *Proceedings of the 22th International Joint Conference on Artificial Intelligence*, 1866-1871, 2011.

[Nastase, 2008] V. Nastase. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP'08*), 763-772, 2008.

[Newell, 1990] A. Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.

[Pirolli and Fu, 2003] P. Pirolli and W. Fu. 2003. Snif-act: A model of information foraging on the world wide web. In *Proceedings of the 9th International Conference on User Modeling*, 45-54, 2003.

[Radev *et al*., 2002] D. R. Radev, E. Hovy, K. McKeown. Introduction to the Special Issue on Summarization. *Computational Linguistics*. 28(4), 399-408. 2002.

[Reitter and Lebiere, 2012] D. Reitter and C. Lebiere. Social Cognition: Memory Decay and Adaptive Information Filtering for Robust Information Maintenance, In *Proceedings of 26th AAAI Conference on Artificial Intelligence*, 242-248, 2012.

[Salton and Buckley, 1988] G. Salton and C. Buckley. Weighting approaches in automatic text retrieval. *Information Processing and Management*. 24(5), 513–523, 1988.

[Singhal, 2001] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24, 35-43, 2001.

[Wu *et al*., 2008] H. C. Wu, R. W. P. Luk, Kam and K. L. Kwok. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26(3), June 2008.

[Xu and Zhuge, 2012a] B. Xu and H. Zhuge. Faceted navigation through keyword interaction. *World Wide Web Journal*, November 2012. DOI:10.1007/s11280-012-0192-2.

[Xu and Zhuge, 2012b] B. Xu and H. Zhuge. Automatic faceted navigation. *Future Generation Computer Systems*, December 2012. DOI:10.1016/j.future.2012.12.003.

[Zhuge, 2010] H. Zhuge. Interactive semantics. *Artificial Intelligence*. 174(2), 190-204, 2010.

[Zhuge， 2011] H. Zhuge, Semantic linking through spaces for cyber-physical-socio intelligence: A methodology, *Artificial Intelligence*, 175, 988-1019, 2011.

[Zhuge, 2012] H. Zhuge, The Knowledge Grid: Toward Cyber-Physical Society, World Scientific, 2012, 2nd Ed.