The Model of Words Cumulative Influence in a Text

Dmytro Lande[1,2](✉), Andrei Snarskii[1,2], and Dmytro Manko[1]

[1] Institute for Information Recording, NAS of Ukraine, Kyiv, Ukraine
dwlande@gmail.com
[2] Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine

Abstract. A new approach to evaluation of the influence of words in a text is considered. An analytical model of the influence of words is presented. The appearance of a word is revealed as a surge of influence that extends to the subsequent text as part of the approach. Effects of individual words are accumulated. Computer simulation of the spread of influence of words is carried out; a new method of visualization is proposed as well. The proposed approach is demonstrated using an example of J.R.R. Tolkien's novel "The Hobbit." The proposed and implemented visualization method generalizes already existing methods of visualizing of the unevenness of words presence in a text.

Keywords: Influence of words · Word weight · Visualization
Computer modeling · Impact model · Analytical model

1 Purpose

The influence of a unit of a text (a word in particular) on the entire text is a hot topic in the task of practicing natural language processing. This problem can be solved within the framework of the theory of information retrieval by determining "weights" of words. At the same time, influence of words on the perception of a subject reading the text with such a memory that the words having been read are forgotten, or cleared out form the memory with time, has not been ever taken into account. This work is devoted to this task. It is the solution of this problem that this work is devoted to, and the corresponding model is described below. Fragments of a text, the total weight of the influence of words, in which the highest one can be considered as the most important for subsequent analytical processing, are provided as well.

2 Approaches to Weighting Words

In the theory of information retrieval, the most common ranking of words weights is performed by the criterion of Salton TF IDF [1], where TF (Term Frequency) is the frequency of occurrence of a word within the selected document, and IDF (Inverse Document Frequency) is a function (most often, a logarithmic function) of the value, inverse to the number of documents, in which the word appeared:

O. Chertov et al. (Eds.): ICDSIAI 2018, AISC 836, pp. 249–256, 2019.
https://doi.org/10.1007/978-3-319-97885-7_25

$$w_i = tf_i \cdot \log\frac{N}{n_i}, \tag{1}$$

where w_i is the weight of word i, tf_i is the frequency of word t_i in a document, n_i is the number of documents in the information array, in which the word is used, N is the total number of documents in the information array. At present, there are many modifications of this algorithm, the most famous of which is Okapi BM25 [2].

Evaluation of the unevenness of the occurrence of words is also possible on the basis of purely statistical variance estimates. In work [3], such an estimate of the discriminant power of the word is given as follows:

$$\sigma_i = \frac{\sqrt{\langle d^2\rangle - \langle d\rangle^2}}{\langle d\rangle}, \tag{2}$$

where $\langle d\rangle$ is the mean sequence value $d_1, d_2, \ldots, d_n$, n is the number of occurrences of a word t_i in the information array. If we denote the coordinates (numbers) of occurrence of a word t_i in the information array by $e_1, e_2, \ldots, e_n$, then $d_k = e_{k+1} - e_k$ $(e_o = 0)$.

Another way for determining weighted meanings of words is proposed in [4]. The idea of the work is based on the concept of graphs of horizontal visibility, in which nodes correspond not only to numerical meanings, but to words themselves. A network of words using the horizontal visibility algorithm is built in three stages. First, a number of nodes is marked on the horizontal axis, each of which corresponds to words in the order of appearance in the text, and weighted numerical estimates are plotted along the vertical axis. At the second stage, a traditional graph of horizontal visibility is constructed [5]. There is a link between the nodes, if they are in "line of sight," i.e. if they can be connected by a horizontal line that does not intersect any vertical line. Finally, the network obtained at the previous stage is compactified. All nodes with a given word are combined into one node (naturally, the index and the position number of the word disappear in this case). All links of such nodes are also combined. It is important to note that no more than one connection remains between any two nodes; multiple links are withdrawn. The result is a new network of words is a compactified graph of horizontal visibility (CHVG). The normalized degree of a node corresponding to a particular word is ascribed to the weight of the word.

3 Visualization of Words Occurrence in a Text

In order to visualize the uneven occurrence of words in the texts, in [6], the technology of spectrograms was proposed, which outwardly resembles the barcodes of goods [3]. But at the same time, this method doesn't allow for considering occurrences of words in different scales of measurements in comparison wavelet analysis.

An algorithm and several examples of visualization of word occurrences are shown in [3], depending on the width of the observation window, which we applied to the fragment of J.R.R. Tolkien's novel "The Hobbit" (Fig. 1). The spectrogram shows the numbers of occurrences of the analyzed word in the text (starting with value 1 at the very bottom) versus the width of the observation windows (occurrence of the word in this case is highlighted in light gray). If several target words occurred in the corresponding monitoring window, then it is covered darker. Expert linguist in appearance can immediately determine the degree of uniformity of occurrence in the text of the analyzed word.

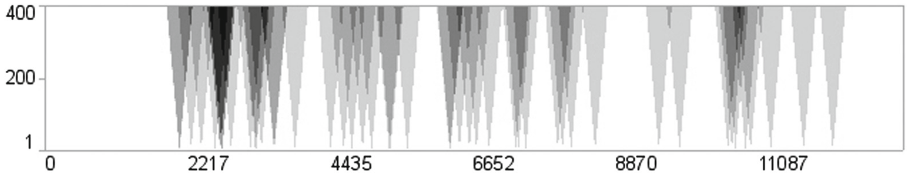

Fig. 1. Spectrogram for the lexeme "Gandalf" in the starting fragment of J.R.R. Tolkien's novel "The Hobbit"

4 The Influence Model of Words

The strength of the influence of words, V, in accordance with the proposed model increases dramatically when a word appears in the text (a point x_i), and then gradually decreases, which is set by a special parameter b. In addition, the magnitude of the influence is determined by the weights σ_i of words and determined by methods considered above. This allows one to consider this model as an extension of well-known "weight" models. The following analytical expression is proposed for determining the strength of the influence of words:

$$V(x,x_i) = \begin{cases} \sigma_i \dfrac{(x-x_i)}{b^2} \exp\left(-\dfrac{(x-x_i)}{b}\right), & \mathit{if}\ (x-x_i) > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where x is the point at which the influence of a word is calculated, x_i is the word appearance position, σ_i is the weight of a word, b is the "memory" parameter of a word.

The code of the MATLAB program for calculating the influence of a single word is given as follows:

```
b=0.8
x=0:0.01:20
T=1
f=((x-T)/(b*b)).*exp(-(x-T)/b);
g=(sign(f)+1).*f/2;
plot(g)
```

Figure 2 shows two peaks realizing the formula (3) with different values of the parameter b.

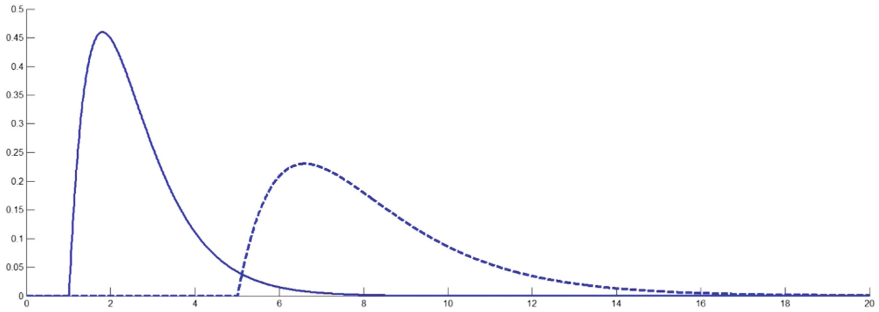

Fig. 2. Dependence of changes in word weights in accordance with (1) with different values of *b* (0.8—solid line, 1.6—dashed line)

We can assume that the power of influence of individual words does not add up, and the influence of a new word absorbs the influence of previous words. In the proposed model, it is assumed, however, that the force of influence of individual words accumulates, summing up in accordance with the following formula:

$$SV(x, X) = \sum_{x_i \in X} V(x, x_i) \tag{4}$$

The code of the MATLAB program for calculating and visualizing the cumulative effect of words in the text model for a fixed parameter *b* is as follows:

```
b=0.8
x=0:0.01:20
hold on
Slova=[1,3,8,10,14];
g=x-x;
for T=Slova
  f=((x-T)/(b*b)).*exp(-(x-T)/b);
  g=g+(sign(f)+1).*f/2;
end
plot(g)
```

Figure 3 shows the result of the “accumulation” of bursts realizing the formula (4) for a fixed value of the parameter *b*.

Calculation of the accumulative effect of words in the text model and its visualization with the spectrum of the parameter values in the form of the projection of the surface onto the plane (Fig. 4) is performed using MATLAB as follows:

```
clear
[x,b]=meshgrid(0:0.01:20,0.5:0.01:2);
Slova=[1,3,8,10,14]
f=(x)./(b.*b).*exp(-(x)./b);
g=f-f;
for T=Slova
    f=(x-T)./(b.*b).*exp(-(x-T)./b);
    g=g+(sign(f)+1).*f/2;
end
pcolor(g)
```

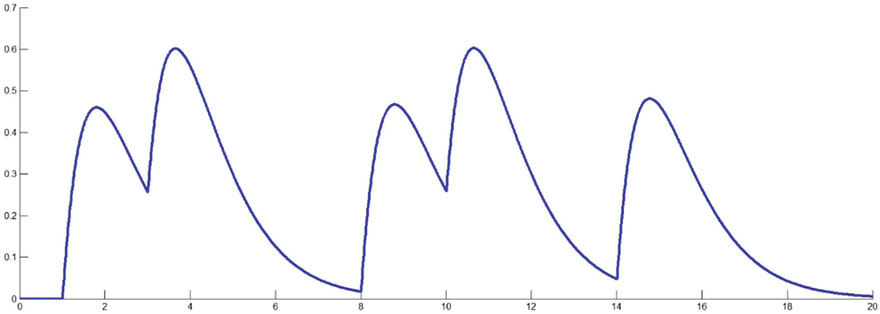

Fig. 3. Graph of accumulative influence of words in the text model

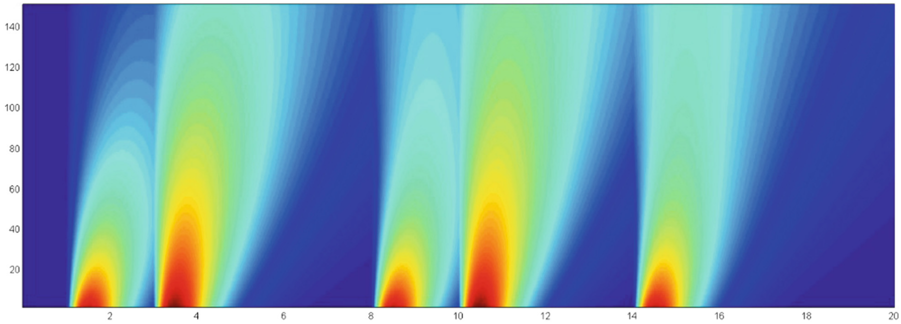

Fig. 4. Number of words in the text versus values of the parameter *b*

Figure 4 represents the accumulative influence of words taking into account the memory parameter in the text model. The shades of the diagram correspond to the value of the memory effect of words.

5 Example

As an example, J.R.R. Tolkien's novel "The Hobbit" was analyzed. In this case, using the CHVG method, the following reference words with weights from 0.9 to 0.6 were identified: BILBO, GANDALF, THORIN, GOBLINS, DWARVES, MOUNTAIN, DOOR, DRAGON, FOREST, ELVES, GOLLUM.

In order to visualize the power of influence of these words, a fragment of the novel was considered (the first 10,000 words). In this fragment, about 300 occurrences of these words were identified. A dependence of their accumulative influence was plotted for the case of a fixed value of $b = 0.1$ (Fig. 5).

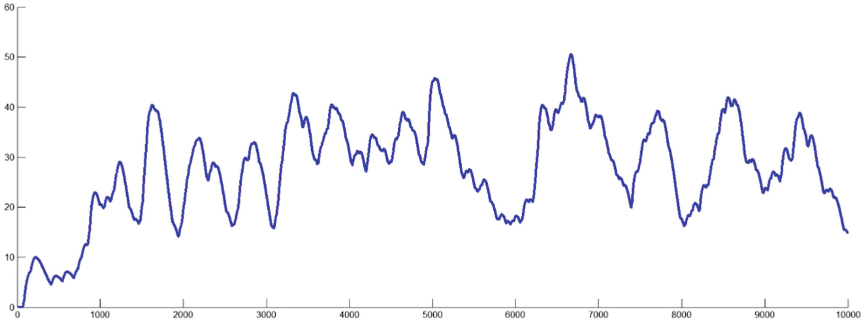

Fig. 5. The accumulative influence of words in J.R.R. Tolkien's novel "The Hobbit"

In Fig. 6, a diagram of the cumulative effect of these words is presented, taking into account the memory parameter b in the range from 0.01 to 1.

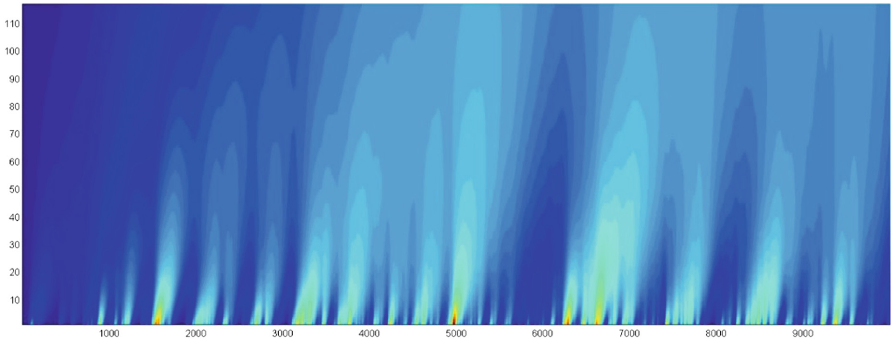

Fig. 6. The accumulative influence of words, taking into account the memory parameter in J.R. R. Tolkien's novel "The Hobbit"

The fragments of the text of the story, corresponding to the peak values in the above diagram, were analyzed.

These fragments of the highest influence of the reduced reference words can be used for the formation of short summaries of the text, as "reference" fragments in analytical and search systems in particular.

Let's reveal such fragments. Peak values: (1600–1700 words). Considered fragment is:

> "Yes, yes, my dear sir—and I do know your name, Mr. Bilbo Baggins. And you do know my name, though you don't remember that I belong to it. I am Gandalf, and Gandalf means me! To think that I should have lived to be good-morninged by Belladonna Took's son, as if I was selling buttons at the door!". "Gandalf, Gandalf! Good gracious me!…"

Peak values: (6600–6900 words). Considered fragment is:

> "A long time before that, if I know anything about the loads East," interrupted Gandalf. "We might go from there up along the River Running," went on Thorin taking no notice, "and so to the ruins of Dale-the old town in the valley there, under the shadow of the Mountain. But we none of us liked the idea of the Front Gate. The river runs right out of it through the great cliff at the South of the Mountain, and out of it comes the dragon too-far too often, unless he has changed."

6 Conclusions

In this paper, a new approach to the evaluation of the influence of words in texts is proposed, an analytical model of the influence of words is presented. The appearance of a word is viewed as a surge of influence that extends to all subsequent text in the proposed model. In this case, the effects of individual words are accumulated. Also, the visualization method presented in the paper, generalizes the existing methods of visualizing the unevenness of the occurrence of words in a text.

The proposed approach can be used to identify a fragment of a text, which is closely related to the most rated words; thereby the most accurately reflects its semantic structure. This allows generating short summaries of texts and digests based on news reports, snippets for search engines, etc.

It should also be noted that in the work, the word is considered as a unit of meaning in the text. If we consider not a single text, but the flow of news messages, and consider an event as a unit, then the proposed model can be generalized for the analysis of text streams in content monitoring systems without changing the mathematical formalism.

References

1. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval, p. 448. McGraw-Hill, New York (1983)
2. Lv, Y., Zhai, C.X.: Lower-bounding term frequency normalization. In: Proceedings of CIKM 2011, pp. 7–16 (2011)
3. Ortuño, M., Carpena, P., Bernaola, P., Muñoz, E., Somoza, A.M.: Keyword detection in natural languages and DNA. Europhys. Lett. **57**, 759–764 (2002)

4. Lande, D.V., Snarskii, A.A.: Compactified HVG for the language network. In: International Conference on Intelligent Information Systems: The Conference is Dedicated to the 50th Anniversary of the Institute of Mathematics and Computer Science, 20–23 August 2013, Chisinau, Moldova, Proceedings IIS/Institute of Mathematics and Computer Science, pp. 108–113 (2013)
5. Luque, B., Lacasa, L., Ballesteros, F., Luque, J.: Horizontal visibility graphs: exact results for random time series. Phys. Rev. E **80**, 046103-1–046103-11 (2009)
6. Yagunova, E., Lande, D.: Dynamic frequency features as the basis for the structural description of diverse linguistic objects. In: CEUR Workshop Proceedings. Proceedings of the 14th All-Russian Scientific Conference “Digital Libraries: Advanced Methods and Technologies, Digital Collections”, 15–18 October, Pereslavl-Zalessky, Russia, pp. 150–159 (2012)
7. Chui, C.K.: An Introduction to Wavelets, p. 366. Academic Press, Cambridge (1992)