УДК 681.3: 004.7

Компактифицированный горизонтальный граф видимости для сети слов

Ландэ Д.В., д.т.н., с.н.с.., Снарский А. А., д.ф.-м.н., проф.
Институт проблем регистрации информации НАН Украины, г. Киев,
Национальный технический университет Украины «КПИ», г. Киев
dwlande@gmail.com, asnarskii@gmail.com

Предлагается метод построения компактифицированного горизонтального графа видимости для сети слов. Определено, что получаемые таким образом сети являются безмасштабными, а также, что среди узлов с наибольшими степенями имеются слова, определяющие не только структуру связности текста, но и его информационную структуру.

Введение

Построение сетей, узлами которых являются элементы текста, слова или словосочетания, фрагменты естественного языка, в некоторых случаях позволяет выявлять структурные элементы текста, без которых он теряет свою связную структуру, информационно-значимые элементы, а также второстепенные для понимания текста слова. Такие сети могут использоваться также для идентификации таких нетрадиционных компонент текста, как колокации, сверхфразовые единства [1], для нахождения подобных фрагментов разных текстов [2].

Существует множество подходов к построению сетей из текстов, так называемых сетей слов (language network), и различные способы интерпретации узлов и связей, что приводит, соответственно, к различным представлениям таких сетей. Узлы могут быть соединены между собой, если соответствующие им слова стоят рядом в тексте [3, 4], принадлежат одному предложению [5], соединены синтаксически [6, 7] или семантически [8, 9].

На стыке теорий цифровой обработки сигналов (Digital Signal Processing) и теории сложных сетей существует несколько методов построения сетей из значений временных рядов, среди которых можно назвать множество методов построения графов видимости (см. обзор [10]), в частности, так называемый горизонтальный граф видимости (Horizontal Visibility Graph -HVG) [11,12]. Базируясь на этих подходах из текстов, в которых отдельным словам или словосочетаниям некоторым образом

поставлены в соответствие числовые значения, также можно строить сетевые структуры. В качестве функции, ставящей в соответствие слову число, можно, например, рассматривать порядковый номер уникального слова в тексте, длину слова, «веса» слов в текстах, например, общепринятую оценку TFIDF (в каноническом виде, равную произведению частоты слова в фрагменте текста – term frequency – на двоичный логарифм от величины, обратной количеству фрагментов текста, в которых это слово встретилось – inverse document frequency) или ее варианты [13, 14], другие весовые оценки слова в тексте.

Сети горизонтальной видимости

При построении сетей слов в данной работе выбрана дисперсионная оценка важности слов [15]. Если пронумеровать все слова в тексте из N слов подряд ($n=1,...N$, n - порядковый номер слова в тексте - позиция слова), то расположение некоторого слова, например A, можно обозначить как $A_k(n)$, где индекс $k=1,2,...,K$ означает номер появления данного слова в тесте, а n - позиция данного слова в тексте. Например, $A_3(50)$ означает, что на 50-й позиции текста находится слово A, которое встретилось третий раз.

Интервал между последовательными появлениями слова при таких обозначениях будет величина $\Delta A_k = A_{k+1}(m) - A_k(n) = m - n$, где на m-м и n-м позициях в тесте находится слово A, которое встретилось $k+1$-й и k-й разы.

Предложенная в [16] дисперсионная оценка рассчитывается как

$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle},$$

где: $\langle \Delta A \rangle$ – среднее значение последовательности $\Delta A_1, \Delta A_2, ..., \Delta A_K$, $\langle \Delta A^2 \rangle$ – последовательности $\Delta A_1^2, \Delta A_2^2, ..., \Delta A_K^2$, K – количество появления слова A в тексте.

В отличие от остальных рядов, изучаемых в рамках цифровой обработки сигналов, ряды из цифровых значений, соответствующих словам, преобразуются в графы горизонтальной видимости, в которых узлам соответствуют не только цифровые значения, но сами слова.

Сеть слов с использованием алгоритма горизонтальной видимости строится в два этапа. На первом этапе строится традиционный граф гори-

зонтальной видимости [16]. Для этого на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются дисперсионная оценка (визуально – набор вертикальных линий, см. рис.1). Между узлами существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. Этот (геометрический) критерий можно записать, согласно [10,11] следующим образом: два узла (слова) слова, например, $B_3(n)$ и $C_7(m=n+5)$ соединены связью, если, см. рис. 1, $\sigma_n, \sigma_m > \sigma_p$, для всех $n < p < m$.

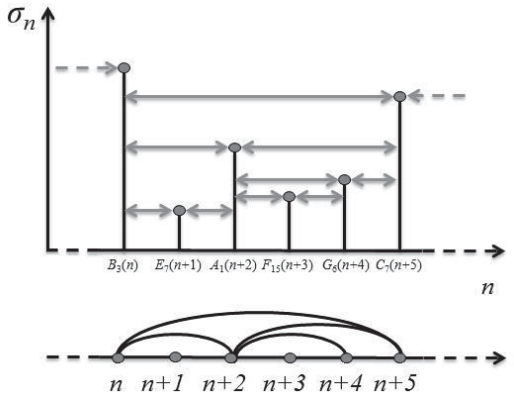


Рис. 1. Пример построения графа горизонтальной видимости

Алгоритм построения КГГВ

Алгоритм построения можно представить удобным для вычисления способом. Так например, на рис. 1 для узла-слова $A_1(n+2)$ смежными в сети считаются слова $B_3(n)$ и $C_1(n+5)$ (и устанавливаются ребра-связи), такие что $B_3(n)$ – ближайшее слева от $A_1(n+2)$ слово, с дисперсионной оценкой $\sigma_n = \sigma_B$, превышающей дисперсионную оценку слова A $\sigma_{n+2} = \sigma_A$, а $C_7(m=n+5)$ – ближайшее справа от $A_1(n+2)$ слово, для которого $\sigma_{105} > \sigma_{102}$.

На втором этапе, полученная на первом этапе сеть компактифицируется. Все узлы с данным словом, например словом A, объединяются в

один узел (естественно, индекс и номер положения слова при этом исчезают). Все связи таких узлов также объединяются. Важно отметить, что между узлами при этом существует не более одной связи, кратные связи изымаются – см. рис. 2.

В частности это означает, что степень (число связей) узла A не превышает суммы степеней $\sum_k A_k(n)$. В результате получается новая сеть слов – *компактифицированный горизонтальный граф видимости* (КГГВ) – рис.2.

В качестве текстов при построении КГГВ рассматривались роман романа М. Булгакова «Мастер и Маргарита», роман Г.Мелвилла «Моби Дик» и массивы новостной информации из веб-пространства.

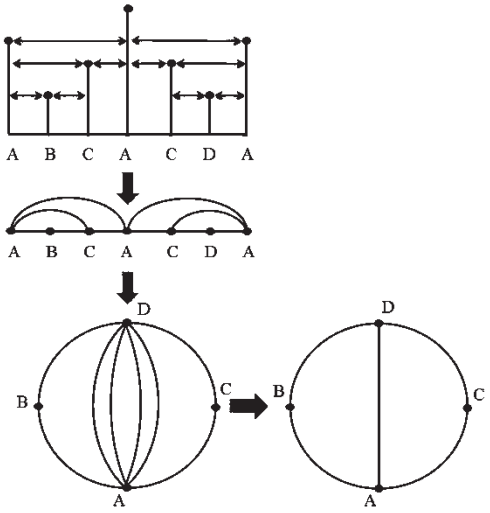
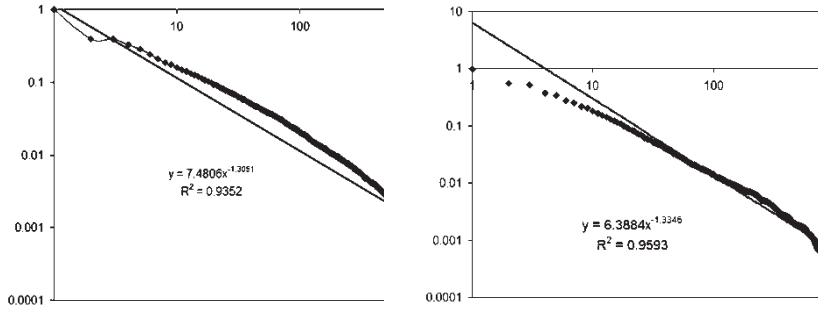


Рис. 2. Два этапа построения компактификационного графа горизонтальной видимости

Некоторые результаты

Для всех построенных КГГВ-сетей слов было определено распределение степеней узлов, которое оказалось близким к степенному (рис. 3), т.е. эти сети являются безмасштабными.



а) б)

Рис. 3. Распределение степеней узлов (в логарифмической шкале) для КГГВ, соответствующих романам «Мастер и Маргарита» (а) и «Моби Дик, или Белый кит» (б). По горизонтальной оси – степени узлов k, по вертикальной – значения 1 – F(k), где F(k) – функция распределения степеней узлов

В состав узлов с наибольшими степенями в этом случае, наряду с личными местоимениями и другими служебными словами (частицы, предлоги, союзы и т.д.), попали слова, определяющие информационную структуру текста [16, 17].

Для сравнения исследованы характеристики простейших сетей языка, когда не первом этапе построения сети связываются соседние слова, входящие в текст, а на втором происходит компактификация сети. Очевидно, вес узлов в этой сети соответствует частоте встречаемости слов, а их распределение – закону Ципфа [18]. При этом самые большие степени имеют узлы, соответствующие словам с наибольшей частотой – союзам, предлогами и т.п., имеющим большое значение для связности текста, но малоинтересным с точки зрения информационной структуры.

Если обозначить Ψ – множество из N различных слов (рассматривался случай $N = 100$), соответствующих наиболее весомым узлам приведенной простейшей сети языка, а Λ – множество из слов, соответствующих наиболее весомым узлам КГГВ, то множество $\Omega = \Lambda \setminus \Psi$ соответствует информативным словам, имеющим, кроме того, важное значение и для связности текста. В Приложении приведены сопоставления 100 наиболее весомых узлов (топ-100) для двух рассматриваемых типов сетей слов по роману М. Булгакова «Мастер и Маргарита» и Г. Мелвилла «Моби Дик, или Белый кит».

В частности, в КГГВ-сети, соответствующей роману М. Булгакова «Мастер и Маргарита» в состав множества Ω попали такие слова, как Иван, Мастер, Варенуха, Берлиоз, Бегемот, Римский, профессор, Левий, Иешуа. В состав множества Ω по роману Г. Мелвилла «Моби Дик, или Белый кит» попали слова: whale, Ahab, boat, Queequeg, sperm, Starbuck, Pequod.

Выводы

В результате исследований сетей слов получены результаты:

Предложен алгорим построения компактифицированного горизонтального графа видимости (КГГВ) на снове последовательности дисперсионных оценок слов текста.

Алгоритм апробирован на различных текстах. Оказалось, что для литературных текстов среди узлов, соответствующих КГГВ с наибольшими степенями, присутствуют слова, не только обеспечивающие связность структуры текста, но и определяющие его информационную структуру, отражающие семантику литературных произведений.

Литература

1. *Солганик Г. Я.* Синтаксическая стилистика. Сложное синтаксическое целое. – 2-е изд., испр. и доп. – М.: Высш. шк., 1911. – 182 с.
2. *Broder A.* Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching, 2000. – P. 1-10.
3. *Ferrer-i-Cancho R., Sole R. V.* The small world of human language // Proc. R. Soc. Lond, 2001. – B 268, 2261.
4. *Dorogovtsev S.N., Mendes J. F. F.* Language as an evolving word web // Proc. R. Soc. Lond, 2001. – B 268, 2603.
5. *Caldeira S. M. G., Petit Lobao T. C., Andrade R. F. S., Neme A., Miranda J. G. V.* The network of concepts in written texts // The European Physical Journal B, volume 49 issue 4 (31 January 2006). – P. 523-529.
6. *Ferrer-i-Cancho R., Sole R.V., Kohler R.* Patterns in syntactic dependency networks // Phys. Rev. E 69, 051915 (2004).
7. *Ferrer-i-Cancho R.* The variation of Zipf's law in human language. // Phys. Rev. E 70, 056135 (2005).
8. *Motter A. E., de Moura A. P. S., Lai Y.-C., Dasgupta P.* Topology of the conceptual network of language // Phys. Rev. E 65, 065102(R) (2002).
9. *Sigman M., Cecchi G. A.* Global Properties of the Wordnet Lexicon // Proc. Natl. Acad. Sci. USA, 99, 1742 (2002).

9. *Nunez A. M., Lacasa L., Gomez J. P., Luque B.* Visibility algorithms: A short review // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, ch. 6., 2012. – P. 119 – 152.
10. *Luque B., Lacasa L., Ballesteros F., Luque J.* Horizontal visibility graphs: Exact results for random time series // Physical Review E, 2009. – P. 046103-1–046103-11.
11. *Gutin G., Mansour T., Severini S.* A characterization of horizontal visibility graphs and combinatoris on words // Physica A, 2011. – 390 – P. 2421-2428.
12. *Jones K.S.* A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation, 1972. – 28 (1). – P. 11–21.
13. *Salton G., McGill M. J.* Introduction to Modern Information Retrieval. – New York: McGraw-Hill, 1983. – 448 p.
14. *Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M.* Keyword detection in natural languages and DNA // Europhys. Lett, – 57(5), 2002. – P. 759-764.
15. *Черняховская Л.А.* Смысловая структура текста и ее единицы // Вопросы языкознания, 1983. – № 6. – С. 118–126.
16. *Giora R.* Segmentation and Segment Cohesion: On the Thematic Organization of the Text // Text. An Interdisciplinary Journal for the Study of Discourse Amsterdam, 1983. – 3. – № 2. – P. 155-181.
17. *Zipf G.K.* Human Behavior and the Principle of Least Effort. – Cambridge, MA: Addison-Wesley Press, 1949. – 573 p.