УДК 004.822

СПОСОБ РАСЧЕТА ВЕСОВЫХ КОЭФФИЦИЕНТОВ ВЕРШИН СЕМАНТИЧЕСКОЙ СЕТИ НАУЧНОГО ТЕКСТА

Аюшеева Н.Н., Кушеева Т.Н.

Восточно-Сибирский государственный университет технологий и управления, Улан-Удэ, e-mail: donir@rambler.ru

В статье представлен способ расчета весовых коэффициентов вершин семантической сети научного текста. Построение семантической сети выполняется для решения задачи автоматического извлечения знаний из текстовых источников. Способ позволяет учитывать основные критерии значимости терминов научного текста, которые являются вершинами семантической сети. К основным критериям отнесены частота встречаемости термина в документе, категория фрагмента текста, в который входит термин, содержательно-смысловой блок, в котором термин встречается. При определении влияния каждого критерия использованы эвристические методы, методы статистической обработки, методы нечеткой логики. Предложенный способ расчета прошел экспериментальную проверку, результаты которой оказались достаточно правдоподобными. Вычисленные по найденной формуле весовые коэффициенты вершин семантической сети действительно адекватно отражают значимость терминов научного текста для определения его смысла.

Ключевые слова: семантическая сеть, весовой коэффициент вершины, научный текст, значимость термина

METHOD OF CALCULATION OF WEIGHT FACTORS TOPS SEMANTIC NETWORK SCIENTIFIC TEXT

Ayusheeva N.N., Kusheeva T.N.

East Siberia State University of Technology and Management, Ulan-Ude, e-mail: donir@rambler.ru

The article shows how to calculate the weights of vertices of a semantic network of scientific text. Building a semantic network is performed to solve the problem of automatic extraction of knowledge from text sources. The method allows to take into account the basic criteria of the scientific significance of the terms of the text, which are the vertices of a semantic network. The main criteria assigned frequency of occurrence of the term in the document, the category of a piece of text, which includes the term semantic content-block in which the term occurs. In determining the impact of each criterion used heuristic methods, statistical treatment, methods of fuzzy logic. The proposed method of calculation was pilot-tested, results of which were quite believable. Calculated from the formula found by weighting the vertices of a semantic network does adequately reflect the significance of the terms of the scientific text to determine its meaning.

Keywords: semantic networks, weighting factor peaks, scientific text, the significance of the term

Одной из задач, решаемых при построении семантической сети текстового документа, является задача определения значимости терминов текста, которые влияют на определение его смысла. При решении данной задачи необходимо учесть множество факторов, влияние которых на значимость терминов различно и не всегда можно определить закономерности этого влияния. Под понятием «значимость» будем понимать, во-первых, «наличие значения, смысла», и, во-вторых, как принято в семиотике и языкознании, отношение знака к другим знакам в рамках языковой системы [1]. Определение значимости неразрывно связано с критериями значимости, роль которых сводится к обнаружению и установлению самого факта наличия значения или смысла термина, который здесь играет роль знака. К основным критериям значимости можно отнести:

– частоту встречаемости термина в документе: чем чаще встречается термин в документе, тем больше отношений он образует с другими терминами;

– категорию текста, в которую входит термин: термины тематической цепочки текста будут более значимы, чем термины текстовой модальности;

– содержательно-смысловой блок, в котором термин встречается: термин, который встретился в основном блоке, будет более полезен для отражения смысла, чем термин, который встретился во вспомогательном блоке.

Для количественного представления значимости терминов обычно используются весовые коэффициенты. *Весовой коэффициент* – числовой коэффициент, параметр, отражающий значимость, относительную важность, «вес» данного фактора, показателя в сравнении с другими факторами, оказывающими влияние на изучаемый процесс [2]. Вычислению весовых коэффициентов терминов предшествуют:

а) оценка степени влияния фактора, который характеризует каждый критерий;

б) определение интегрального показателя весового коэффициента термина.

Рассмотрим определение степени влияния фактора, характеризующего каждый из вышеназванных критериев, на весовой коэффициент термина.

Частота встречаемости термина в документе. Статистический показатель тер-

мина документа невозможно использовать без предварительной обработки. Это связано с тем, что значение частоты встречаемости термина, который чаще других был употреблен в документе, абсолютно не влияет на его значимость. Бóльшее значение будет иметь ранг частоты, который позволяет уравнивать значимости самых встречаемых терминов любых текстов и одновременно распределяет значимости терминов внутри одного текста. При этом термины с одинаковой частотой встречаемости, имея одинаковый ранг частоты, будут одинаково значимы для передачи смысла. Для учета частоты встречаемости при определении весового коэффициента термина предлагаем использовать формулу (1).

$$w_1 = 1 - \log_{\max(r)} r, \qquad (1)$$

где r – ранг частоты термина.

Она позволяет получить нормализованное значение w_1 за счет вычисления логарифмической функции с основанием, равным максимальному рангу частоты. Вычитание из единицы позволяет терминам с наибольшим рангом частоты иметь большее значение w_1, а для терминов с максимальным рангом, т.е. которые редко используются в тексте, этот показатель будет равен нулю, что означает его неважность для отражения смысла текста.

Категория текста. Это одна из существенных характеристик текста, представляющая собой отражение определенной части общетекстового смысла различными языковыми, речевыми и собственно текстовыми (композитивными) средствами. Категория текста имеет знаковую природу, план содержания такого знака – это единый текстовой смысл (например, целостность, тема, тональность, пространство, проспекция), а план выражения – функционально ориентированная типовая композиция разноуровневых языковых средств [3].

Категория текста является единицей анализа, несущей в себе основные свойства целого, а именно целенаправленность и композитивность. Каждая текстовая категория воплощает в себе отдельную смысловую линию текста, выраженную группой языковых средств, особым образом организованной в относительную внутритекстовую целостность. Совокупность категорий текста, дополняющих друг друга и переплетающихся между собой, создают текст в качестве коммуникативной системы.

В соответствии с категориально-текстовой концепцией, основанной на принципе отражательности, категория текста как смысловая часть текста отражает один из компонентов коммуникативного акта, в число которых входит предмет речи; субъект(-ы) речи, то есть автор(-ы) текста в целом; оценочная точка зрения субъекта; его эмоционально-психологический настрой; пространство и время как неотъемлемые атрибуты ситуации, в которой порождается текст; адресат общения. Соответственно выделяются текстовые категории темы, субъекта (авторизации), оценочности, тональности (текстовой модальности), текстового пространства и времени, адресата. В силу объективно дробного выражения каждой категории в тексте к ним добавляется структурная текстовая категория композиции. На наш взгляд, именно текстовые категории темы и композиции являются в настоящий момент наиболее важными для определения значимости терминов, и, кроме того, более прозрачными для исследования.

Тема – существенный и необходимый признак всякого текста [4]. Это экстралингвистический фактор, который входит в ядро текста и определяет его структуру. Тема выражается в тематических группах, которые составляют тематическое поле тематического единства. Тематическую группу научного текста, в частности научной статьи, можно сформировать, выделив термины из заголовка и подзаголовков. При этом если частота встречаемости выделенных терминов будет высокой в тексте, то их с полной уверенностью можно включить в текстовую категорию темы. Тогда вклад в значение весового коэффициента термина можно принять равным 1, если термин отражает тему текста, и 0 в противном случае:

$$w_2 = \begin{cases} 1, \text{ термин отражает тему} \\ 0, \text{ в противном случае} \end{cases} \qquad (2)$$

Говоря о второй текстовой категории, выбранной в работе, то композиция текста представляет собой единство внутренней структуры содержания, внешнего его деления на части и сами эти части. Для выделения такой структуры можно использовать выделение формальных текстовых признаков.

Содержательно-смысловой блок. Текстовая категория композиции соотносится с понятием *содержательно-смысловой блок*. Научный текст состоит из логически выделенных содержательных блоков: блок постановки и понимания проблемы (Проблема), блок изучения опыта предшественников (Опыт), блок изложения варианта решения проблемы, доказательства и аргументов (Решение), блок обобщения полученных данных и подведения итогов (Итог). Для идентификации каждого блока применяется метод выделения формальных

текстовых признаков, которые с высокой вероятностью используются в конкретном блоке. Кроме вышеперечисленных блоков в текстах можно выделить, так называемые, дополнительные блоки, которые играют большую роль для отражения коммуникативной, аспектной, семантической, информативной, функционально-смысловой структуры научного текста: для описания общеизвестного и доказанного факта (Факт); для выражения убежденности автора (Убежденность); для обеспечения межфразовой связи (Коннектор); для отражения информации, противоположной претексту (Противоположность); для отражения информации о часто/редко повторяющихся событиях (Повторяемость); для отражения развития информации (Развитие); для уточнения информации (Уточнение). Для достаточно небольших текстов, которыми являются научные статьи, наличие дополнительных блоков является не характерным: некоторые блоки могут отсутствовать, некоторые блоки могут быть очень маленькими и содержать в себе только 1–2 термина. В связи с этим на данном этапе работы будут рассмотрены термины четырех основных блоков: Проблема, Опыт, Решение, Итог. Очевидно, что для передачи основного замысла научной статьи существенную роль играют блоки Проблема, Решение и Итог. При этом блок Решение составляет зачастую большую половину текста. Это видно по результатам исследования корпуса научных статей объемом 100 единиц по различным областям знаний [6]. На этот же факт указывает существующее большое число маркеров и индикаторов, характерных для рассматриваемого блока. В связи со сказанным весовые коэффициенты терминов блока примем равными согласно (3).

$$w_2 = \begin{cases} 0{,}30, & \text{если термин блока Проблема,} \\ 0{,}15, & \text{если термин блока Опыт,} \\ 0{,}25, & \text{если термин блока Решение,} \\ 0{,}30, & \text{если термин блока Итог.} \end{cases} \quad (3)$$

Исследование научных текстов статей позволило выделить наиболее характерные индикаторы и маркеры каждого содержательно-смыслового блока. Если термин используется в предложении, содержащем формальный признак того или иного блока, то его вес корректируется на соответствующую величину. При этом если термин встретился в более, чем одном блоке, его вес изменяется на сумму соответствующих величин. Частота встречаемости термина в пределах одного блока здесь не играет роли, поскольку этот показатель был учтен в формуле (1).

Вычисление интегрального весового коэффициента термина. Бесспорно, что вышеуказанные три критерия значимости термина по-разному влияют на значение его весового коэффициента. Тогда интегральный весовой коэффициент может быть рассчитан по формуле.

$$W = \sum_{i=1}^{m} k_i w_i, \quad (4)$$

где k_i – весовой коэффициент критерия i, $i = 1..3$.

Для определения весовых коэффициентов критериев воспользуемся процедурой взвешивания, предложенной в работе [5]. Для реализации данной процедуры необходимо осуществить две операции: вычислить критериальные индексы q_i, на основе которых затем определяются весовые коэффициенты k_i для всех критериев. Исходной информацией для определения степени важности каждого критерия значимости термина служит следующая вопросная конструкция: насколько важен i-й критерий для определения значимости термина текста? Формат возможных ответов может быть представлен следующим множеством:

1) достаточно важен;
2) скорее важен, чем не важен;
3) скорее не важен, чем важен;
4) совершенно не важен;
5) затрудняюсь ответить.

Тогда индекс важности каждого критерия может быть вычислен по формуле.

$$q_i = \left(\frac{1}{1 - y_{in}/100} \sum_{i=1}^{n-1} a_i y_{ij} \right)^{1-py_{in}}, \quad (5)$$

где i – индекс критерия; j – индекс варианта ответа респондентов на вопрос относительно важности i-го критерия; n – общее число предусмотренных вариантов ответа на вопрос (в нашем случае 5); y_{ij} – доля респондентов (в процентах), указавших j-й вариант ответа для i-го критерия; a_i – весовой коэффициент j-го варианта ответа (для всех критериев используется единая шкала весовых коэффициентов; $0 \le a_i \le 1$); p – нормирующий коэффициент, величина которого определяется в ходе вычислительных экспериментов. Для показателя a система весовых коэффициентов для всех критериев одинакова: $a_1 = 1{,}0$; $a_2 = 0{,}6$; $a_3 = 0{,}4$; $a_4 = 0$. Их значения интерпретируются как степени принадлежности рассматриваемого критерия к нечеткому множеству «важный критерий для определения значимости термина текста».

Идентификация индексов (5) позволяет установить иерархию критериев. Для последующего включения всех критериев в интегральный весовой коэффициент необходимо от величин q_i перейти к весовым коэффициентам важности каждого критерия, которые вычисляются по формуле.

$$k_i = \frac{q_i}{\sum_{i=1}^{m} q_i}, \qquad (6)$$

где m – общее число критериев.

Процедура (6) позволяет пронормировать все критерии таким образом, что выполняется классическое балансовое условие.

$$\sum_{i=1}^{m} k_i = 1. \qquad (7)$$

Имея оценки критериальных весовых коэффициентов w_i и коэффициентов их важности k_i, можно рассчитать интегральный весовой коэффициент значимости термина W.

Определение коэффициентов важности критериев. В рамках выполнения вычислительных экспериментов была составлена анкета, включающая вопросы:

1. Насколько важен критерий «Частота встречаемости термина в научном тексте» для определения значимости термина текста?

2. Насколько важен критерий «Термин отражает тему научного текста» для определения значимости термина текста?

3. Насколько важен критерий «Содержательно-смысловой блок» для определения значимости термина текста?

Формат возможных ответов был представлен выше.

Рассчитанные по формуле (5) индексы важности каждого критерия соответственно равны $q_1 = 0{,}675$; $q_2 = 0{,}887$; $q_3 = 0{,}625$.

Вычислив по формуле (6) весовые коэффициенты важности критериев, находим $k_1 = 0{,}309$; $k_2 = 0{,}406$; $k_3 = 0{,}285$.

Вычислительные эксперименты. Рассмотрим на примере взвешивание терминов семантической сети научного текста. Для этого выберем одну статью «Технология многомерных баз данных» из коллекции статей по предметной области «Базы данных». На рисунке приведен фрагмент семантической сети рассматриваемого текста.

Фрагмент семантической сети

Для терминов данного фрагмента в табл. 1 приведены частота их встречаемости, ранг частоты и весовые коэффициенты w_1 первого критерия значимости, рассчитанные по формуле (1). В последнем столбце этой таблицы приведены весовые коэффициенты w_2 второго критерия значимости термина, определенные по формуле (2).

Для расчета весового коэффициента третьего критерия значимости необходимо идентифицировать содержательно-смысловые блоки текста по формальным текстовым признакам, характерным для каждого блока. Будем выделять предложения с характерными индикаторами и маркерами, и в отношении терминов этих предложений будут задаваться весовые коэффициенты по третьему критерию в соответствии с формулой (3), представленные в последнем столбце табл. 1.

По формуле (4) с учетом вычисленных весовых коэффициентов важности критериев $k_1 = 0{,}309$; $k_2 = 0{,}406$; $k_3 = 0{,}285$ находим значения интегральных весовых коэффициентов терминов (табл. 2).

Таблица 1

Характеристики и весовые коэффициенты терминов

Термин	Частота встречаемости	Ранг частоты	Весовые коэффициенты w_1	Весовые коэффициенты w_2	Весовые коэффициенты w_3
Данные	31	1	1,000	0	0,30
Многомерные базы данных	9	5	0,373	1	0,30
Анализ данных	7	7	0,241	0	0,25
Многомерные кубы	6	8	0,189	1	0,25
Многомерная модель данных	4	10	0,102	0	0,15
Область применения	3	11	0,065	0	0,30
Проблематика	1	13	0,000	0	0,30

Таблица 2

Результаты определения интегральных весовых коэффициентов W

Термин	Весовые коэффициенты W
Многомерные базы данных	0,6068
Многомерные кубы	0,5357
Данные	0,3945
Анализ данных	0,1457
Область применения	0,1056
Проблематика	0,0855
Многомерная модель данных	0,0743

Данные в табл. 2 отсортированы по убыванию значений интегральных весовых коэффициентов. Термин «Многомерные базы данных», имея пятый ранг частоты, встречаясь в названии статьи и относясь к содержательно-смысловому блоку «Итог», получил наибольшее значение весового коэффициента, что было ожидаемо. Анализируя другие термины, мы также видим проявление закономерностей, соответствующих выдвинутым предположениям.

Заключение

Полученные результаты являются достаточно правдоподобными и отражают значимость терминов научного текста для определения его смысла. В результате выполненной работы предлагается модифицировать формулу (5), так как в рамках данной задачи можно пренебречь степенью $(1 - py_{in})$, поскольку погрешность вычисления, оцененная в сотых долях, вполне приемлема и практически не влияет на результат дальнейших вычислений. В перспективе требуется обосновать выбор системы весовых коэффициентов вариантов ответов при обработке результатов экспертного опроса, применяя методы нечеткой логики.

Список литературы

1. Уфимцева А. А. Знаковые теории языка // Лингвистический энциклопедический словарь / под ред. В.Н. Ярцевой. – М.: Советская энциклопедия, 1990.

2. Райзберг Б. А. Современный экономический словарь / Б.А. Райзберг, Л.Ю. Лозовский, Е.Б. Стародубцева. – 5-е изд., перераб. и доп. – М.: ИНФРА-М, 2007. – 495 с.

3. Стилистический энциклопедический словарь русского языка / под ред. М.Н. Кожиной. – М.: Флинта, Наука, 2003.

4. Русский язык: учебно-методический ресурс для студентов и школьников // ГОУ ВПО «Российский государственный профессионально-педагогический университет». – http://rulinguistic.com.

5. Балацкий Е. В. Методы диагностики социального самочувствия населения // Мониторинг общественного мнения. – 2005. – №3.

6. Найханова Л.В. Роль формальных текстовых признаков для построения семантической сети научного текста / Л.В. Найханова, Н.Н. Аюшеева, Т.Н. Кушеева // Вестник ВСГТУ. – 2011. – №1(32). – С. 32–37.

References

1. Ufimtseva A.A. Sign language theory – Encyclopedic dictionary of linguistics. Moscow, Soviet encyclopedia, 1990.

2. Rayzberg B.A., Lozovskij L.UJ., Starodubtseva E.B. Modern economic dictionary. Moscow, INFRA-M, 2007. 495 p.

3. Stylistic encyclopedic dictionary the Russian language. By: Kozhina, Margarita Nikolaevna. Moscow, Flinta Publ., 2003.

4. Russian language. Educational-methodical resource for students and pupils. Available at: http://rulinguistic.com.

5. Balatskiy E.V. Methods of diagnostics of social well-being of the population – Monitoring of public opinion, 2005, no. 3

6. Naykhanova L.V., Ayusheeva N.N., Kusheeva T.N. Role of formal text signs at construction of the semantic network of the scientific text – Bulletin of ESSUTM, 2011, no.1. 32-37 p.

Рецензенты:

Найханова Л.В., д.т.н, профессор, заведующий кафедрой систем информатики Восточно-Сибирского государственного университета технологий и управления, г. Улан-Удэ;

Ширапов Д.-Д.Ш., д.ф.-м.н., профессор, заведующий кафедрой электронно-вычислительных систем Восточно-Сибирского государственного университета технологий и управления, г. Улан-Удэ.

Работа поступила в редакцию 05.06.2012.