УДК 004.9

МЕТОД КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ВАЖНОСТИ СЛОВОСОЧЕТАНИЙ В СИСТЕМАХ СЕМАНТИЧЕСКОГО АНАЛИЗА ЭЛЕКТРОННЫХ ТЕКСТОВ

Чалая Лариса, Чижевский Антон, Шевякова Юлия

Харьковский национальный университет радиоэлектроники, Украина

Аннотация

Рассматривается метод количественной оценки важности словосочетаний в семантической сети онтологической базы знаний. Метод основан на комплексном использовании значений показателей TF, TF/IDF, а также общего и частных рангов слов. Исходной операцией предложенного алгоритма является предварительное упорядочение слов по убыванию важности и составление модифицированных списков. Экспериментально исследовано влияние длины словосочетаний, а также количества слов из первого и второго модифицированных списков на вероятность присутствия анализируемых словосочетаний в документах корпуса текстов.

Consider a method quantitative estimation of the importance of word combinations in the semantic network of ontological knowledge base. The method is based on the integrated use of values of TF, TF/IDF, as well as general and special grades of words. The initial operation of the proposed algorithm is pre-ordering the words in descending order of importance and preparation of the modied list. The eect of the length of word combinations, as well as the number of words from the rst and second modied list to the probability of the presence word combinations in the documents of the analyzed corpus.

Введение

К важным задачам создания онтологий следует отнести определение наиболее важных слов для каждого отдельного текста, а также для всего корпуса текстов в целом. Наиболее важными словами, отображающими суть предметной области, можно считать слова, которые наиболее часто встречаются как в каждом отдельном тексте, так и в возможно большем количестве текстов из анализируемого корпуса текстов.

Корпоративная онтологическая база знаний представляет собой, как правило, совокупность разного рода слабо структурированных документов, в которых часто присутствуют, т.е. некоторые ситуации и решения, которые уже были ранее приняты в аналогичных ситуациях. В сетях, которые используют такие базы знаний, поиск решения заключается в поиске в этих базах наиболее подходящих прецедентов и соответствующих им документов. Эффективность поиска решений в базах знаний прецедентов в значительной мере зависит от используемых методов поиска. Современные поисковые системы основаны в основном на применении полнотекстового поиска – поиска в каждом из документов всех терминов, входящих в запрос. При этом для оценки релевантности (важности) документов учитываются частота встречаемости слов в документе и их средняя языковая частотность.

В настоящее время наибольшее распространение для оценки важности слов в онтологиях получили показатели TF (term frequency) и IDF (inverse document frequency). Показатель TF определяется как отношение числа вхождения в документ некоторого слова к общему количеству слов документа [1]. Показатель IDF соответствует инвертированному значению частоты, с которой некоторое слово встречается в документах коллекции. Учёт IDF уменьшает вес широкоупотребительных слов. Эффективным методом поиска специфичных слов, характерных для анализируемого документа, является применение комбинированного коэффициента TF/IDF [2]. По таким показателям оценивается важность слова в пределах отдельного документа или анализируемого корпуса текстов. В то же время представляется целесообразным решить задачу выделения в тексте не только отдельных слов, но и наиболее важных словосочетаний, отражающих семантику текста.

Описание метода количественной оценки важности словосочетаний

В данном докладе предлагается метод поиска таких словосочетаний, основанный на количественной оценке важности элементов текста с использованием значений TF, TF/IDF и так называемых рангов слов. Исходной операцией здесь является предварительное упорядочение по убыванию важности и составление соответствующих списков важных слов для показателей TF и TF/IDF. Под частными рангами слова R_1 и R_2 будем понимать значения величин, обратных номерам позиции этого слова в упорядоченных списках для TF и TF/IDF соответственно.

Под общим рангом слова R будем понимать коэффициент, который соответствует наибольшему из значений частных рангов (R_1 и R_2) этого слова для анализируемого корпуса текстов. После определения значений R_1 и R_2 производим модификацию исходных списков. При этом слово относим к первому модифицированному списку, если для него $R_1 > R_2$, и, соответственно, ко второму, если $R_2 > R_1$.

Очевидно, что важность словосочетаний для текущего текста зависит от наличия в них слов каждого из модифицированных списков. Например, словосочетание, полностью составленное из слов первого списка, недостаточно хорошо отображает семантику текста, т.к. все слова в нем, скорее всего, окажутся слишком общими; с другой стороны, если словосочетание составлено полностью из слов второго списка, в нем могут отсутствовать ключевые слова, имеющие максимальную частоту вхождения в документы анализируемого корпуса текстов.

Важно также правильно выбрать максимальную длину оцениваемых словосочетаний. Очевидно, что с ее увеличением уменьшается вероятность присутствия в тексте осмысленных словосочетаний заданной длины.

В связи с этим представляется целесообразным в общий критерий количественной оценки важности словосочетаний ввести нормированный коэффициент $K(W_i)$, который будет зависеть как от количества

вхождений в словосочетание W_i слов из первого и второго модифицированных списков, так и от длины словосочетания.

Анализ списков словосочетаний (для представительного набора текстов по направлению «Компьютерные науки») позволил экспериментально оценить влияние длины словосочетаний, а также количества слов из первого и второго модифицированных списков на вероятность присутствия этих словосочетаний в документах анализируемого корпуса текстов. Очевидно, что такая вероятность может быть принята в качестве коэффициента $K(W_i)$, значения которого находятся в диапазоне [0; 1].

В таблице 1 представлены значения коэффициента $K(W_i)$ для разных типов словосочетаний.

Предлагается оценивать важность словосочетания, перемножая ранги отдельно взятых слов. При этом, поскольку ранги слова нормированы от нуля до единицы, то с увеличением длины словосочетания уменьшается результат такого произведения. В связи с этим его необходимо умножить на количество слов в рассматриваемом словосочетании.

Следует также учитывать, что слова, для которых $R_1 > R_2$, определяют специфику анализируемого текста, то есть слова из первого списка с большой вероятностью попадут в большинство словосочетаний, характерных для этого текста. Следовательно, оценки важности таких словосочетаний будут близки. Наиболее же специфичными для конкретного текста являются слова второго списка, для которых $R_2 > R_1$, следовательно, именно за счет этого можно повысить релевантность критерия оценки важности словосочетания.

Таблиця 1 – Значения коэффициента $K(W_i)$ для разных типов словосочетаний

№	Количество слов в словосочетании по первому списку	Количество слов в словосочетании по второму списку	$K(W_i)$
1	2	0	0,3
2	1	1	1
3	3	0	0,2
4	2	1	0,6
5	1	2	0,7
6	4	0	0,1
7	3	1	0,5
8	2	2	0,8
9	1	3	0,8

Кроме того, величина модуля разности рангов слов по разным спискам ($_1j - R_2j$ |) влияет на вероятность того, что слово, которому соответствует максимальный ранг, является важным для анализируемого корпуса текстов: чем больше минимальный модуль разности двух рангов одного и того же слова, тем выше находятся все слова из словосочетания в каком-либо из двух списков и тем важнее словосочетание в целом.

Таким образом, для комплексной оценки важности словосочетания W_i в рассматриваемом тексте можно предложить следующий коэффициент $M(W_i)$

$$M(W_i) = K(W_i) * \left(\prod_{i=1}^{n} R_i\right) * min(|R_1j - R_2j|) * n * max\left(\frac{TF}{IDF_j}\right), j = 1, \cdots, n; R_1j \neq R_2j,$$

где n – количество слов в словосочетании W_i (не считая стоп-слов);

$K(W_i)$ - коэффициент из таблицы 1;

$R; R_1; R_2$ – общий и частные ранги слова соответственно.

Таким образом, общий алгоритм поиска наиболее важных для текста словосочетаний состоит в следующем:

Шаг 1. Предварительно ранжируются по важности исходные списки слов по заданному корпусу электронных текстов (по R_1 и R_2) и определяются их общие ранги.

Шаг 2. Производится модификация ранжированных списков по приведенным выше правилам.

Шаг 3. Формируется исходный список словосочетаний из слов модифицированных списков.

Шаг 4. Определяется оценка важности словосочетаний $M(W_i)$.

Выводы

Предложенный метод может использоваться для задач семантического поиска в системах анализа электронных текстов и автоматического создания онтологий.

Список использованных источников:

1. Чижевский, А. В. Комбинированный критерий оценки важности слов в формируемых онтологиях [Текст] / А.В. Чижевский, Ю.Ю. Шевякова // Радиоэлектроника и молодежь в XXI веке: материалы 16-го Междунар. молодежного форума, 17–19 апр. 2012 г. – Х. : ХНУРЭ, 2012. – С. 55 – 56.
2. TF-IDF [Электронный ресурс] / Wikipedia. – Режим доступа: http://ru.wikipedia.org/wiki/TF-IDF