Раковская Елена Евгеньевна

аспирант

ФГБОУ ВПО «Байкальский государственный

университет экономики и права»

г. Иркутск, Иркутская область

ВЕКТОРНАЯ МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ

***Аннотация**: в современных условиях большое внимание уделяется математической обработке и моделированию текстовой информации. В статье в доступной и понятной форме изложены основные теоретические представления векторной модели текстов, рассмотрены возможности определения «веса» терминов, модифицирования показателей частот, что может быть полезным во многих прикладных задачах – в информационном поиске, категоризации текстовой информации и рубрикации документов.*

***Ключевые слова**: модели представления текстов, векторная модель, «взвешивание» терминов.*

В настоящее время наблюдается лавинообразный рост массивов информации, сконцентрированной в корпорациях, учреждениях, государственных департаментах, управлениях социального обеспечения и т. д., а также в глобальной сети Интернет. Очень часто данная информация представлена в текстовом виде, не структурирована и не приведена к единому формату. Невозможно пренебрегать такого рода массивами данных, следовательно, разработка методов и моделей обработки текстовой информации – актуальная задача на современном этапе [2].

В большинстве прикладных программ, использующих естественно-языковые тексты, применяется векторная модель [4; 5]. Теоретическое представление векторной модели состоит в следующем:

Имеется корпус D с N документами и словарь V с M терминами. Представление векторного пространства документов определяется как элемент n-мерного векторного пространства R^M[1]:

$$\vec{d_i} = \left(w_i^{[1]}, w_i^{[2]}, w_i^{[3]}, \ldots, w_i^{[M]}\right) \in \mathbb{R}^M$$

где $w_i^{[j]}$ показывает вес j-того термина t_j в словаре для документа d_i.

Для выбора веса термина в документе в классическом варианте возможны следующие характеристики:

1. По частоте термина определяется, как часто представлен определенный термин в специальном документе.

2. Частота документов указывает, как много существует документов в корпусе, в которых встречается термин. Частота документов интерпретируется как глобальный фактор веса.

Если термин представлен во многих, или почти во всех документах, то «важность» его незначительна. Это свойство моделируется посредством обратной (инверсной) частоты документов:

$$w_{\text{global}}(t_j) = idf(t_j) = \frac{N}{df(t_j)}$$

где N – количество всех документов в корпусе, $df(d_j)$ – частота документов с термином t_i. Для выравнивания значений инверсной частоты документа часто применяется логарифмирование:

$$w'_{\text{global}}(t_j) = \log\left(idf(t_j)\right) = \log\left(\frac{N}{df(t_j)}\right)$$

Логарифмирование уменьшает вес очень редких терминов, одновременно веса очень частых терминов становятся меньше. На рис. 1. мы видим два графика, в которых значение веса представлены с логарифмированием и без него – для рис. 1 (b) отчетливо идентифицируется сглаживающее развитие логарифмического взвешивания.

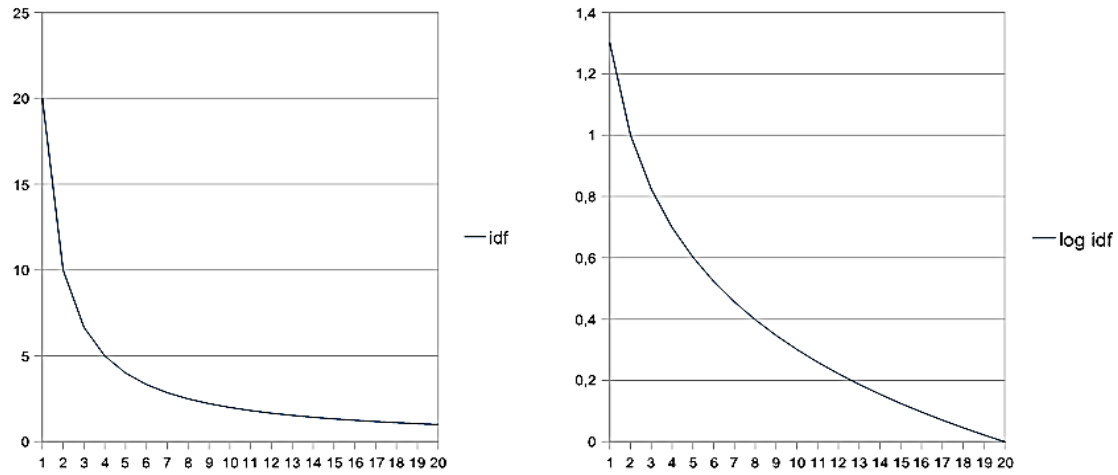

Рис. 1. Глобальный вес для нормальной (a) и логарифмической (b) обратной частоты документа

Глобальное взвешивание выступает, как можно сказать уже из названия, не как особенность одного специального документа. Следовательно, должен учитываться еще локальный вес. Важность термина внутри определенного документа можно рассчитать посредством частоты термина. Чем чаще появляется термин в документе, тем важнее он оценивается для этого документа. Появляется термин редко, только один раз, или даже вовсе не появляется, тогда он вносит малый вклад, или вообще не вносит вклад в содержание.

Как локальный вес принимается частота $tf_{d_i}(t_j)$ термина t_j в документе d_i:

$$w_{\text{lokal}}(t_j, d_i) = tf_{d_i}(t_j)$$

Комбинация из приведенных локальных и глобальных взвешиваний известна в ИП (информационном поиске) под названием TF-IDF. Посредством перемножения обоих весов получается общий вес термина в документе:

$$w_{\text{TF-IDF}}(t_j, d_i) = w_{\text{lokal}}(t_j, d_i) \cdot w'_{\text{global}}(t_j) = tf_{d_i}(t_j) \cdot \log\left(\frac{N}{df(t_j)}\right)$$

Отсюда вытекает такая особенность, что вес термина, который не представлен в документе, или представлен во всех документах, равен 0.

В целом, определено векторное представление для документа. Документ d_i представляется как вектор d_i, в котором элемент массива каждого термина по показателю TF-IDF составляет:

$$\vec{d_i} = (w_{\text{TF-IDF}}(t_1, d_i), w_{\text{TF-IDF}}(t_2, d_i), w_{\text{TF-IDF}}(t_3, d_i), \ldots, w_{\text{TF-IDF}}(t_M, d_i))$$

Следует отметить, что векторная модель – это далеко не единственная модель представления текстовой информации [3]. Наиболее известны булева модель, вероятностная модель. В булевой модели документы моделируются как совокупность терминов. Или термин содержится в документе, или не содержится в нем. Имеется также расширенные булевы модели, в которых определяются различия важности термина для документа с применением ранжирования и функций сходства.

Список литературы

1. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко. – СПб.: БХВ-Петербург, 2007. – 384 с.

2. Леонтьева Н.Н. Автоматическое понимание текстов. Системы, модели, ресурсы: Учеб. пособие / Н.Н. Леонтьева. – М.: Academa, 2006. – 304 с.

3. Маннинг К.Д. Введение в информационный поиск. / К.Д. Маннинг, П. Рагхаван, Х. Шютце. – М. – СПб. – Киев: Вильямс, 2011. – 520 с.

4. Сэлтон Г. Автоматическая обработка, хранение и поиск информации / Г. Сэлтон. – М.: Сов. Радио, 1973. – 560 с.

5. Salton G. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Information Processing & Management. – 1988. – №5 (24). – P. 513–523.