ОБ ОДНОМ ПОДХОДЕ К ОЦЕНКЕ ИНФОРМАЦИОННОЙ ЗНАЧИМОСТИ РЕЗУЛЬТАТОВ ПОИСКА ПУТЕМ СЕМАНТИЧЕСКОЙ ОПТИМИЗАЦИИ ЗАПРОСА

Д.В. Гринченков, к.т.н., доцент, Д.Н. Кущий, ассистент
Южно-Российский государственный политехнический университет (НПИ)
имени М.И. Платова,
г. Новочеркасск
E-mail: grindv@yandex.ru, dkushchiy@rambler.ru

Одним из основополагающих моментов поиска электронных образовательных ресурсов с целью обновления методического обеспечения является обработка текста рабочей программы учебной дисциплины с формированием эталонной выборки и построением информационного запроса, являющегося аналогом стандартного запроса пользователя на естественном языке [1].

Для оценки эффективности с позиции содержательной значимости необходимо решить практическую задачу семантической оптимизации запроса, заключающуюся в построении шаблонов коллокаций в контексте рассматриваемой дисциплины [2].

Пусть $S^{\varepsilon} = \left\{ s_1^{\varepsilon}, s_2^{\varepsilon}, ..., s_n^{\varepsilon} \right\}$ – множество всех значимых слов рабочей программы в лемматизированном виде, полученное в результате предобработки содержания рабочей программы путем удаления стоп-слов и общезначимых словосочетаний с помощью разработанных словарей и критерия t-score[3]. Для определения терминологических словосочетаний предметной области будем рассчитывать вес связи между рассматриваемыми понятиями с последующим отсечением низких значений:

$$w\left(s_i^{\varepsilon}, s_j^{\varepsilon}\right) = \frac{nr\left(w_1 + w_2\right)}{NR} + \frac{NR}{MR} + f\left(s_i^{\varepsilon}, s_j^{\varepsilon}\right)\left(\frac{1}{f\left(s_i^{\varepsilon}\right)} + \frac{1}{f\left(s_j^{\varepsilon}\right)}\right), \quad (1)$$

где *nr* – число встречаемости в тексте данной пары понятий с данным отношением,
NR – общее число встречаемости данного отношения в тексте,
MR – общее число отношений в тексте,
w_1 – вес первого понятия,
w_2 – вес второго понятия,
$f\left(s_i^{\varepsilon}, s_j^{\varepsilon}\right)$ – частота совместной встречаемости этих двух понятий по любому отношению в тексте,
$f\left(s_i^{\varepsilon}\right)$ – частота встречаемости первого понятия в тексте,
$f\left(s_j^{\varepsilon}\right)$ – частота встречаемости второго понятия в тексте.

В современных моделей текстовых документов коллокации часто рассматриваются как пары слов, встречающиеся рядом друг с другом. Для повышения точности будем рассматривать коллокаты, которые могут находиться на некотором расстоянии. Его будем обозначать $k_{i,j} = \overline{1,l},\ l \in \left\{ Z | l \le l\_max \right\}$, *l_max* – максимальное расстояние между термами.

Информационный запрос *q* на основе эталонной выборки примет вид:

$$q = \sum_{i,j}^{n} s_i^{\varepsilon}\, s_j^{\varepsilon} \left\{ k_{i,j} \right\}. \quad (2)$$

Описанное выражение будет использовано для оценки степени удовлетворения информационной потребности и ранжирования результатов. Обозначим множество термов сопоставляемого с эталонной выборкой документа как $S^{\tau} = \left\{ s_1^{\tau}, s_2^{\tau}, \ldots, s_m^{\tau} \right\}$.

При оценке сходства текстов следует учитывать множество пар коллокаций словоупотреблений $\Theta\left(S^{\varepsilon}, S^{\tau}\right)$, в которых первый элемент пары входит в состав эталонного текста, а второй элемент входит в сопоставляемый текст, и данные словоупотребления совпадают по нормальной форме терма s^d:

$$\Theta\left(S^{\varepsilon}, S^{\tau}\right)\left\{\left\langle s_i^{\varepsilon}\ s_j^{\varepsilon},\ s_i^{\tau}\ s_j^{\tau}\right\rangle \mid s_i^{\varepsilon}, s_j^{\varepsilon} \in S^{\varepsilon} \exists s_i^{\tau}, s_j^{\tau} \in S^{\tau} : \left\langle s_i^{\varepsilon}, s^d \right\rangle, \left\langle s_i^{\tau}, s^d \right\rangle \wedge \left\langle s_j^{\varepsilon}, s^d \right\rangle, \left\langle s_j^{\tau}, s^d \right\rangle\right\}.$$

Данное выражение позволяет обеспечить полноту и сохранить релевантные результаты выборка информации учётом омонимов нормальных форм.

Расчет общей информационной значимости содержания сопоставляемого текста можно определить следующим выражением:

$$R^{\varepsilon\tau}\left(S^{\varepsilon}, S^{\tau}\right) = \sum_{s_i^{\varepsilon} s_j^{\varepsilon},\ s_i^{\tau} s_j^{\tau} \in \Theta\left(S^{\varepsilon}, S^{\tau}\right)} a v\left(s_i^{\varepsilon} s_j^{\varepsilon}\right) v\left(s_i^{\tau} s_j^{\tau}\right), \qquad (3)$$

где α - коэффициент, характеризующий степень неточности совпадения исходных словоформ,

$v\left(s_i^{\varepsilon}\ s_j^{\varepsilon}\right)$–вес пары словоупотреблений (коллокации).

В качестве функции для определения весов $v\left(s_i^{\varepsilon}\ s_j^{\varepsilon}\right)$используется классический метод tf (termfrequency), в котором вес определяется как функция от количества вхождений терма в документе [4], с ограничением суммы весов словоупотреблений текста-эталона до единицы. Кроме того, в функции поиска должна быть реализована поправка на метаданные документов, выраженных, в том числе, в нетекстовой форме или на естественном языке.

Таким образом, анализ содержания рабочей программы образовательной дисциплины позволит не только определить ассоциативные связи между ключевыми словами, необходимые для оценки пертинентности, но и найти их весовые коэффициенты для уточнения значения алгоритмической релевантности по завершению информационного поиска.

Список литературы

1. Гринченков Д.В., Кущий Д.Н. Актуальность и принципы построения интеллектуальной информационной системы формирования методического обеспечения учебных дисциплин на основе ресурсов сети интернет // Изв. вузов. Сев.-Кавк. регион. Техн. науки – 2014. – № 3. – С. 114-119.

2. Гринченков Д.В., Кущий Д.Н. К вопросу формирования нечетких коллокаций на основании семантического анализа рабочих программ образовательных дисциплин // Проблемы модернизации инженерного образования в России: Сб. науч. статей по проблемам высшей школы / Юж.-Рос. гос. политехн. ун-т (НПИ) -Новочеркасск : ЮРГПУ(НПИ), 2014. - С. 298-300.

3. GriesS.Th. Statistics for Linguistics with R // A Practical Introduction. – Berlin, Boston: De Gruyter, 2013.–354 p.

4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб.пособие/ Большакова Е.И., Клышинский Э.С., ЛандэД.В.,Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. – 272 с.