Статистико-лексикографический подход к индексированию двуязычных текстовых массивов

Ландэ Д.В., Дармохвал А.Т., Жигало В.В.
Информационный центр «ЭЛВИСТИ», Киев, Украина

Проблема индексирования текстовых документов – выделения лексических единиц, совокупность которых может выступать поисковыми образами документов, особенно актуальна в связи с лавинообразным ростом контента сети Интернет. Если для обеспечения традиционного информационного поиска зачастую бывает достаточно построения простого полнотекстового индекса, то для таких задач, как, например, выявление содержательных дубликатов документов, поиск плагиата, построение сюжетных цепочек, качественное автоматическое выделение «опорных» лексических единиц (ЛЕ) играет решающее значение.

Для обеспечения эффективного функционирования системы интеграции и мониторинга новостей InfoStream, авторами предложен подход к индексированию документов, базирующийся как на статистических, так и на лексикографических методах. Задача состояла в индексировании документов из коллекции, представленной на двух языках – русском и украинском. При этом интенсивность пополнения данной коллекции в сутки составляет около 60 тыс. документов, сканируемых роботами системы InfoStream более чем с 2500 украино- и русскоязычных веб-сайтов.

Для индексирования использовались украино- и русскоязычные словарные массивы. Ввиду технической сложности представления полных лексикографических баз данных для двух языков, авторами использовался лишь относительно небольшой, но, по-видимому, самый существенный для данной задачи срез - множество имен существительных, дополненное некоторыми фамилиями, аббревиатурами, названиями компаний. Как показал опыт, такой подход полностью себя оправдал как для обеспечения качества индекса, так и для визуализации результатов работы. В предложенной процедуре индексирования для выделения наиболее значимых ЛЕ использовался статистический метод, базирующийся на применении общеизвестного подхода TF IDF, а точнее его модификации Okapi BM25, в которой каждой ЛЕ t из документа D приписывается вес $W(t,D)$ по формуле:

$$W(t,D)=\frac{f(t,D)(k+1)}{f(t,D)+k(1-b+b|D|/L)}\cdot\log\frac{N-n(t)+0.5}{n(t)+0.5},$$

где $f(t,D)$ - частота встречаемости ЛЕ t в документе D, $|D|$ - длина документа D, L - средняя длина документа в коллекции текстов, общее количество которых - N, $n(t)$ - количество документов в коллекции, содержащих данную ЛЕ, k, b - параметры, выбираемые экспертами. Для «обучения» системы использовалась коллекция документов за 2007 год, объемом свыше 10 млн. документов (новостей, публикуемых на веб-сайтах).

Определенное количество ЛЕ с максимальными весами, присутствующих в соответствующих словарных массивах, нормализовались и попадало в индекс в качестве опорных. Опорные ЛЕ дополнительно переводились, для чего использовались украино-русский и русско-украинские словари, которые обеспечивают перевод от 97 до 99% выделяемых ЛЕ. Для частичного решения проблем омонимии, проявляемой как на этапе экстрагирования опорных ЛЕ, так и на этапе их перевода, использовались статистические критерии – выбирались наиболее частотные значения ЛЕ.

Рассматриваемый подход используется в системе интеграции и мониторинга новостей InfoStream для решения многих задач, серди которых можно назвать построение сюжетных цепочек, дайджестов, информационных портретов, выявление содержательных дубликатов, поиск содержательно подобных документов не зависимо от языка, на котором они представлены.