ТЕХНОЛОГИЯ ПОЛНОТЕКСТОВОГО ПОИСКА В МУЛЬТИЯЗЫЧНЫХ СЕТЕВЫХ РЕСУРСАХ

Д.В. Ландэ,
Институт проблем регистрации информации НАН Украины,
dwl@visti.net

В.В. Жигало,
ООО «Информационный центр «ЭЛВИСТИ»,
vladlen@visti.net

В статье описывается технология построения мультиязычных сюжетных цепочек с использованием опорных слов и их переводов на другой язык. Предложенная технология позволят отображать актуальные сюжетные цепочки по мере добавления новых документов из потока новостных документов, формировать параллельные текстовые корпусы.

На сегодняшний день задача поиска информации в мультиязычных ресурсах очень актуальна. Очевидна ситуация когда человеку нужно получить полную информацию о событии, при этом язык первоисточников отходит на второй план. В частности, проблема поиска в мультиязычных ресурсах является главной для построения параллельных корпусов документов, формирования сюжетных цепочек.

Выявление четкого дублирования информации сегодня не представляет проблем, однако подобные по смыслу сообщения, нечеткие дубликаты находятся не так легко, здесь на помощь приходят специальные алгоритмы [1-2].

На практике явные дубликаты выявляются даже с помощью механизмов контрольных сумм, но такой подход не решает проблем пользователей, для которых чаще всего не имеет значения, с чем они имеют дело, с прямой перепечаткой или с небольшой перефразировкой. Определяющими в этом случае являются такие характеристики как скорость обработки запросов, достоверность отклика (например, оцениваемая по источникам), а также дополнительные сервисы – возможность нахождения документов, подобных уже имеющимся, подключения средств автоматического реферирования и перевода и, конечно же, уточнения запроса.

В настоящее время наряду с другими проблемами в области информационного поиска, очень важными являются проблемы многократного дублирования информации, избытка шумовой информации, спама, а также отсутствие мультиязычных средств поиска.

В данной статье описывается подход, с помощью которого в ходе поисковых процедур определяются нечеткие дубликаты документов, приведенных на разных языках (реализован поиск дубликатов, приведенных на украинском и русском языках). Процедура выявления дубликатов построена на использовании методов извлечения опорных слов на основе эмпирико-статистических свойств текстов с помощью частотного морфологического словаря, а также перевода этих слов на другой язык.

Процедуру выявления дубликатов можно представить в виде нескольких этапов:
- создание морфологических словарей;
- создание частотных словарей - обучение системы;
- создание словарей переводов;
- построение программами поиска опорных слов;
- создание процедур поиска дубликатов на разных языках.

Сначала создаются морфологические словари, которые для каждой словоформы содержат ее нормальную форму. Это нужно для того, чтобы в дальнейшем можно было привести все найденные словоформы к нормальной форме.

Далее создается частотный словарь на базе морфологического словаря, в котором записывается частота каждой словоформы, найденной в процессе «обучения» частотного словаря на тестовом массиве документов.

Для построения электронных морфологических словарей используются имеющаяся электронная версия словаря Зализняка, который насчитывает около 93 тыс. слов в нормальной форме для русского языка, и бесплатный словарь ispell, который насчитывает около 1 миллиона украинских словоформ, соответственно, для украинского языка. Для выявления опорных слов для документов использовались частотные словари для различных языков, для чего использовались доступные морфологические словари и корпус документов, сканируемых из Интернет системой контент-мониторинга InfoStream [3]. Морфологические словари были дополнены известными именами фирм и фамилиями известных личностей.

В частотном словаре для каждого слова записывается количество его появлений в некотором большом массиве документов, а также количество документов, в которых нашлось это слово.

При машинном обучении частотного словаря из каждого документа в корпусе выделялись словоформы, которые приводились к нормальной форме. Для эффективности поиска опорных слов в результирующие словари входили только те слова, которые встретились в корпусе документов более двух раз. Также было решено использовать только имена существительные.

Определение опорных слов основано на использовании подхода *TF IDF* [4], а точнее его модификации Okapi BM25 [5]:

$$TF \cdot IDF = \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \cdot \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

где $f(q_i, D)$ – частота встречаемости слова q_i в документе D, $|D|$ – длина документа D, $avgdl$ - средняя длина документа в коллекции текстов, общее количество которых - N, $n(q)$ – количество документов в коллекции, содержащих данное слово, k_1, b – параметры, выбираемые экспертами.

Для дальнейшей работы с каждого документа выбирались 12 опорных слов с наибольшими значениями *TF IDF*.

Исходные данные для построения словарей переводов были получены путем перевода имен существительных в нормальной форме существующими программами перевода текстов.

В случае если одному слову соответствовало несколько переводов, то выбиралось наиболее употребляемое значение в соответствии частотным словарем.

Происходит поиск нормальной формы для каждой из словоформ. В случае омонимии, выбирается та нормальная форма словоформы, которая наиболее является частотной в словаре. Далее происходит подсчет количества словоформ. Вычисляется опорные слова с помощью формулы Okapi BM25.

После вычисления весовых коэффициентов происходит ранжирование нормальных слов и выбираются самые первые двенадцать. Полученные двенадцать опорных слов переводятся на другой язык с помощью словарей переводов. Все опорные слова и слова переводы приписываются к документу и выдаются в выходной поток.

Экспертные оценки показали, что удалось добиться 99% качества при переводе опорных слов [6].

При выявлении сюжетных цепочек в системе контент-мониторинга InfoStream для каждого документа выполняется определение его опорных слов, а затем поиск похожих документов по 6-ти опорным словам из 12 опорных слов других документов (рис. 1). В случае если находится подобный документ, то он приписывается к сюжетной цепочке,

уже сформированной ранее. Если подобных документов не находится, выполняется формирование новой сюжетной цепочки.

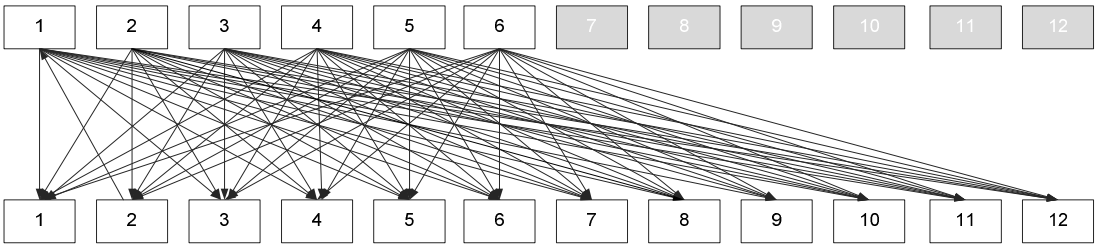

Рис. 1. Сравнение опорных слов

При формировании мультиязычных сюжетных цепочек в случае поступления документов на разных языках (рис. 2) для их связывания используются не только опорные слова на том языке на котором документ описан, но и на других языках, используя переведенные опорные слова. При поиске подобных словах на различных языках также как и в случае формирования сюжетов 6 опорных слов документа сравниваются с переводами 12 опорных словах других документов. Кроме того, берется 6 переведенных опорных слов документа, которые сравниваются с 12-ю опорными словами в документах на другом языке.

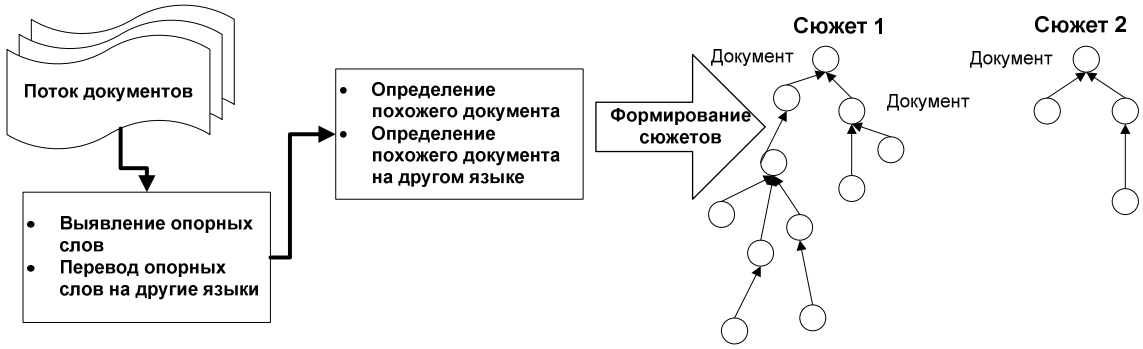

Рис. 2. Схема выявления сюжетных цепочек

Традиционные методы поиска информации в массивах данных не позволяют выявить похожих документов на разных языках а также степень их подобия. В данной статье были показаны компоненты поисковой технологии (поиска документов на разных языках), в частности, методы построения сюжетных цепочек документов. Степень похожести документов в системе характеризуется количеством совпадающих опорных слов документов и их переводов. Данная технология работает в системе контент-мониторинга InfoStream и позволяет отображать наиболее актуальные сюжеты в информационном сегменте, охватываемом данной системой.

Литература

1. Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю. Подход к выявлению дублирования сообщений в новостных информационных потоках // Труды 9 Всерос. научн. конф.: Электронные библиотеки: перспективные методы и технологии, электронные коллекции – RCDL’2006. Суздаль, Россия, 2006. – С. 115-119.

2. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9 Всерос. науч. конф.: Электронные библиотеки:

перспективные методы и технологии, электронные коллекции – RCDL’2007. Переславль-Залесский, Россия, 2007. – С. 166-174.

3. Григорьев А.Н., Ландэ Д.В. Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Компьютерная лингвистика и интеллектуальные технологии: Труды междунарожной конференции "Диалог’2005" (Звенигород, 1-6 июня 2005 г.). – М.: Наука, 2005. – С. 109-111.

4. Salton G, Buckley C. Term-Weighting Approaches // Automatic Text Retrieval. Information Processing and Management, 1988. – № 24, 5. – P. 513-523.

5. Robertson S. E., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond // Foundations and Trends in Information Retrieval, 2009. – V. 3. – № 4. – P. 333–389.

6. Ландэ Д. В., Жигало В. В. Подход к созданию многоязычных параллельных корпусов веб-публикаций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции "Диалог’2009" (Бекасово, 27-31 мая 2009 г.). – Вып. 8 (15). – М.: РГГУ, 2009. – С. 278-283.