Подход к выявлению подмножеств похожих документов

© А.В. Антонов, С.Г. Баглей, В.С. Мешков

Корпорация «Галактика»
{alexa, baglei, meshkov}@galaktika.ru

Аннотация

В статье описывается подход к определению подмножеств похожих документов, реализованный в поисково-аналитической системе Галактика-Zoom. Метод основан на работе со статистико-лингвистическими данными, составляющими Информационный портрет (ИнфоПортрет), формируемый в системе Галактика-Zoom. ИнфоПортрет формируется в соответствие каждому документу в заданном подмножестве документов. Элементами ИнфоПортрета являются языковые инварианты (слова и пары слов), статистически отличающие данную выборку документов. Основная задача, которая решается с помощью описываемого подхода – разбиение заданного множества документов на подмножества похожих документов и подмножества уникальных документов. В качестве критерия близости в подмножествах выбрано расстояние Дженсена–Шеннона.

1 Введение

Общей проблемой, снижающей эффективность работы пользователя с поисковой системой, является избыточность информации при выдаче результатов по запросу. Причинами возникновения избыточности могут являться, например, нечетко сформулированные запросы к поисковой системе, омонимичность элементов поисковых запросов, а также другие.

Задача уменьшения избыточности, может решаться различными способами. Один из наиболее действенных способов снижения избыточности – диалог пользователя с системой, то есть, режим, при котором пользователю предоставляется возможность уточнения своих информационных предпочтений.

Как часть диалога пользователя с системой может рассматриваться формирование подмножеств похожих документов, которое позволяет пользователю рассматривать некоторое подмножество в качестве некоторой отдельной сущности. Работая с полученными подмножествами похожих документов, пользователь может более быстро воспринимать данные, с которыми он работает.

Исследования проводились в рамках моделей, использующих расстояния *Kullback-Leibler* [6] или – *Jensen-Shannon* [5] между двумя распределениями признаков. Описание применяемой в системе модели приведено в работе [4]. Мы использовали этап формирования ИП – множества языковых инвариантов (слов и словосочетаний) для выявления признаков, характеризующих каждый документ из рассматриваемого множества. Выбор наиболее тематически значимых слов для работы метода являлся интуитивным, и основывался на наших предыдущих работах в этой области. Кроме того, мы учитывали необходимость сохранения быстродействия в работе системы. Для этой цели требовался этап предварительной фильтрации признаков-слов и словарных пар, характеризующих отдельный документ. Естественным шагом было использование в качестве смыслового фильтра существующий в системе механизм построения Информационного Портрета документа.

В качестве меры близости документов рассматривается расстояние между распределениями весов элементов (слов и/или словосочетаний) их ИП. Опорным считается ИП наибольшего размера, то есть, такой ИП, в котором количество составляющих его элементов максимально. Для формирования подмножеств документов был выбран порог. Похожими считаются документы, у которых расстояние не меньше заданного.

Построение подмножеств документов осуществляется с помощью утилиты сравнения документов выборки и измерения расстояния между ними.

Галактика-Zoom представляет собой поисково-аналитическую систему обработки больших объемов неструктурированных данных. Подробно архитектура, принципы работы, характеристики системы описаны в работах [1, 2].

Основным объектом в системе Галактика-Zoom является понятие Информационного портрета выборки документов (ИнфоПортрета). ИнфоПортрет представляет собой список языковых инвариантов (слов и словосочетаний), отличающих данную выборку от прочих. Технология построения информационного портрета, детально описанная в работах [2, 3, 4], основана на статистических методах обработки текстовой информации. Используя характеристики элементов сформированного ИнфоПортрета и собственной статистики документа, возможно формирование информационного портрета отдель-

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

ных документов. То есть, для каждого документа система формирует список слов и словосочетаний, статистически отличающих данный документ от прочих в выборке. ИнфоПортрет представляет собой информацию, описывающую содержание документа в целом.

Система позволяет не прибегать к условию равнозначности слов, то есть существует возможность обработки весовых значений элементов, составляющих ИП. Мы использовали следующие преимущества предоставляемые системой с точки зрения задачи поиска похожих документов:

- возможность получения величины относительной значимости слов и словосочетаний для документа;
- возможность упорядочивания значимых слов и словосочетаний в документе исходя из величины их относительной значимости в выборке.

2 Постановка задачи

2.1 Входные данные

$D = \{d_i\}, i = \overline{0, n-1}$ – множество документов, загруженных в базу.

$P = \{p_i\}, i = \overline{0, n' - 1}$ – множество ИП, соответствующих документам из D.

S – вектор номеров документов из множества D, упорядоченный по убыванию размера ИП.

$C = \{C_i\}, i = \overline{0, k}$ – подмножества похожих документов базы.

U_0 – список документов базы, у которых отсутствует ИП.

U – список уникальных документов.

2.2 Промежуточные данные

S^1 – вектор номеров документов из множества D, упорядоченный по убыванию размера соответствующего им ИП.

S^w – рабочий вектор номеров документов, упорядоченный по убыванию размера соответствующего им ИП, к которому применяется алгоритм построения подмножеств похожих документов.

2.3 Выходные данные

$P^1 = \{p_i\}, i = \overline{n, n + m' - 1}$ – множество ИП документов.

S – вектор номеров документов упорядоченный по убыванию размера соответствующего им ИП.

$C^1 = \{C_i^1\}, i = \overline{0, k_1}$ – подмножества похожих документов.

U_0^1 – список документов, у которых отсутствует ИП.

U^1 – список уникальных документов базы.

2.4 Алгоритм построения подмножеств похожих документов (*ППД*) на векторе S^w

Первый документ из вектора порождает первое подмножество похожих документов и сравнивается со всеми последующими, основываясь на расстояниях Kullback-Leibler или Jensen-Shannon, задающих меру подобия ИП (оценку близости документов). Если данная мера подобия для какого-нибудь ИП оказывается выше заданного порогового значения, то текущий документ приписывается первому подмножеству. Сравнение продолжается, пока не исчерпывается вектор документов.

После этого происходит обработка следующего документа, не вошедшего в первое подмножество, с которым последовательно сравниваются все последующие актуальные документы и т.д.

В результате формируется некоторое неизвестное заранее количество подмножеств похожих документов.

1. Инициализация. Построение и последующее сохранение ИП каждого документа из множества D^1. Формирование и сохранение вектора S^1. Документы, у которых отсутствуют ИП, добавляются в список U^0.

2. Множество D^1 разбивается на два непересекающихся подмножества: D_+^1 – документы, размер ИП которых не меньше наибольшего размера ИП в векторе S^0 и D_-^1 – документы с меньшими размерами.

3. Если подмножество D_+^1 не пустое, то формирование вектора S^w из документов множества D_+^1, центроидов из подмножеств C похожих документов и из списка U уникальных документов. Применение алгоритма *ППД* к документам из множества D_+^1 как возможным новым центроидам. Если центроид из C приписывается к новому подмножеству, то для всех документов похожих на него пересчитывается мера близости относительно нового центроида.

4. Если остались уникальные документы из подмножества D_+^1 или подмножество D_-^1 не пустое, то формирование вектора S^w из уникальных документов подмножеств D_+^1, D_-^1 и центроидов подмножеств похожих документов. Применение алгоритма *ППД* к уникальным документам из подмножества D_+^1 и центроидам с возможным пополнением их новыми документами из D_-^1.

5. Если после выполнения шагов 3 и 4 остались уникальные документы из множества D^1, то формирование вектора S^w из *всех* оставшихся уникальных документов. Применение алгоритма *ППД* к каждому уникальному документу из множества D^1.

Примечание. В процессе выполнения шагов 3, 4 и 5 формируются подмножества C^1 и список U^1.

3 Результаты экспериментов

Для оценки качества работы метода нами было проведено исследование на экспериментальной базе документов. В качестве наполнения базы использовался массив, состоящий из нормативно-справочных документов. Моделировалась ситуация проведения разбиения определенной выборки на множества похожих и множество уникальных документов.

Таблица 1. Основные характеристики базы документов

Параметр	*Количество*
Документов в базе	5000
Слов в базе	63100
Словомест в базе	2008235
Словосочетаний в базе	13621
Мест словосочетаний в базе	172383

Рассматривая работу алгоритма поиска похожих документов как часть функциональности поисковой системы Галактика-Zoom, в качестве элементов исходного набора данных были определены документы тестовой выборки из базы, заданной случайным образом.

В качестве модели документа мы также использовали ИнфоПортрет с соответствующими ему характеристиками, формируемый в системе Галактика-Zoom. Далее приведены основные параметры выборки и полученных подмножеств похожих документов.

Таблица 2. Основные параметры полученных подмножеств

Количество подмножеств	31
Количество документов в выборке	523
Количество документов, включенных в подмножества	426
Количество документов, общих для подмножеств	0
Минимальное число документов в подмножестве	3
Максимальное число документов в подмножестве	21
Минимальное число объектов в подмножестве	3
Максимальное число объектов в подмножестве	10

Анализ полученных результатов показал, что в результате работы алгоритма получено достаточно плотное покрытие исходного массива данных непересекающимися подмножествами. Параметры полноты и точности формирования подмножеств вполне приемлемо для эффективного практического использования алгоритма.

4 Заключение

Мы применили алгоритм построения подмножеств похожих документов, основанный на измерении меры близости между отдельными документами выборки. Основная цель применения алгоритма – представление однородной или близкой к однородной части исходных данных в удобном виде для восприятия пользователем системы.

Экспериментальная проверка показала, что получены приемлемые результаты в части полноты и точности формируемых подмножеств. Таким образом, использование метода оказалось оправданным.

Литература

[1] Антонов А. Методы классификации и технология Галактика-Zoom // Международный форум по информации, Москва, ВИНИТИ, 2003. т. 28.

[2] Антонов А, Курзинер Е. Автоматическое выделение предметной области большого необработанного текстового массива // Компьютерная лингвистика и интеллектуальные технологии : Труды Международного семинара Диалог-2002.

[3] Антонов А. Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации // Сб. ВИНИТИ. – 2001, №8.

[4] Антонов А., Мешков В. Современные проблемы поисковых систем и некоторые пути их преодоления (Сер. «Аналитика–Капитал»). – М., 2000.

[5] Fuglede, B., Topsше, F. Jensen-Shannon Divergence and Hilbert space embedding. // University of Copenhagen, Department of Mathematics, http://www.math.ku.dk/~topsoe/ISIT2004JSD.pdf.

[6] Kullback, S., Leibler, R. // On information and sufficiency, Annals of Mathematical Statistics 22:79–86, 1951.

[7] Антонов А., Баглей С., Мешков В., Суханов А. Кластеризация документов с использованием метаинформации // Компьютерная лингвистика и интеллектуальные технологии : Труды Международного семинара Диалог-2006.

An Approach to Detection of Similar Documents Subsets

Alexander Antonov, Stanislav Bagley, Valentin Meshkov

In this paper, we propose an approach to similar documents detection. We consider automatic decomposition of an initial document set into a collection of non-overlapping subsets of similar documents and a subset of independent documents as the result of an approach work. The model underlying the approach uses statistical information about words and pairs of neighboring words in texts. To measure the proximity between documents we use Jenson-Shannon divergence applied to that statistical information. We

give an example of test run of an approach in the paper performed on a collection of legal documents.