Математические структуры и моделирование
2001, вып. 8, с. 132–140

УДК 519.71

КРИТЕРИЙ БЛИЗОСТИ ДОКУМЕНТОВ И КЛАСТЕРИЗАЦИЯ

О.Г. Чанышев

In this article the algorithm for automatic clustering is presented. This algorithm is based on original model of the real text. Automatically extracted «dominant lexems» are using for automatic clustering of non-grupped beforehand sets of documents.

Введение

Общей проблемой области автоматического анализа естественноязыковых текстов (ЕЯ-текстов), к которой относятся задача автоиндексирования и тесно связанная с ней [1, стр. 341] задача автоматической тематической классификации, является проблема понимания текста системой искусственного интеллекта. Современные промышленные автоиндексирующие и автоклассифицирующие системы [2–4] обладают высоким быстродействием, эргономными пользовательскими интерфейсами. Например, в продукт LinguistX компании Inxight Software входят усовершенствованные средства обработки естественного языка от поисковых механизмов до средств распознавания рукописного текста, включая автоматическое реферирование, извлечение информации и морфологический анализ [5]. С целью развития классических методов кластеризации достаточно широко используются искусственные нейронные сети [6]. Однако прогресс в этой области имеет скорее технологический характер, что А.С. Нариньяни констатировал (несколько эмоционально) следующим образом: «Массированное, продолжавшееся несколько десятков лет наступление в области автоматической обработки текста захлебнулось. По отношению исходных планов и надежд оно окончилось достаточно очевидным провалом.» [7]. Так или иначе, принципиальный вопрос о приемлемой теории текста, на которой должны базироваться методы автоматического анализа текста, остается открытым.

Алгоритмы классификации относятся к обширному классу алгоритмов распознавания образов [8, Глава 9], [9, Глава 4], в основе которых лежит гипотеза компактности: «реализации одного и того же образа обычно отражаются в признаковом пространстве в геометрически близкие точки, образуя «компактные» сгустки» [10, стр. 29]. Методы классификации текстов едва ли не исчерпывающе представлены в монографии Дж. Солтона (Gerard Salton) [1, Глава 8]

E-mail: chanysh@iitam.omsk.net.ru
Омский филиал Института математики СО РАН

вплоть до физической идеи определения кластеров «путем коллапсирования пространства с помощью гравитационного притяжения».

Представляемый в настоящей статье метод автоматической кластеризации ЕЯ-текстов основан на «ассоциативной модели реального текста» [12, 13]. С операционной точки зрения, он относится к «порождающим методам классификации по принципу снизу вверх, при котором все объекты первоначально считаются несгруппированными» [1, стр. 242]. Основное отличие от геометрического подхода заключается в принципиальной несимметричности используемой меры тематической близости, поскольку требование «равной похожести» части и целого семантически не представляется естественным.

1. Доминанты, тематическая близость и кластеризация документов

Ассоциативная модель рассматривает текст как задание тотального графа предметной области списками смежностей лексем – предложениями. Текст не нормализуется, и не рассматриваются лексемы, принадлежащие заданному стоп-множеству. Из оставшихся учитываются только «независимые» лексемы, для любой пары которых существуют минимум два предложения, в которых они встречаются отдельно. В качестве меры важности лексемы используется «ассоциативная мощность» (Ψ), совпадающая с частотой (ω) только в случае задания графа бинарными списками смежности. При этом, если для низко- и среднечастотных лексем можно положить $\Psi \approx Const \times \omega$, то для высокочастотных эта зависимость существенно не монотонна (из $\omega_j > \omega_i$ не следует $\Psi_j \geq \Psi_i$). Анализ ранговых распределений Ципфа-Мандельбродта для независимых лексем позволил ввести понятие критического значения ассоциативной мощности для выделения наиболее важных (доминантных) лексем, объем которых не превышает 0.04 от объема словаря текста.

Хорошо известно, что из двух «подзадач» задачи распознавания образов: выбора множества признаков объекта и распознавания на основе выбранного множества – наиболее трудно формализуемой и в этом смысле наиболее сложной является первая [11]. Эксперименты по автоматическому реферированию текстов показали, что независимые лексемы связи адекватно представляют текст, но их слишком много для попарного сравнения каждого документа с каждым. А это предусматривает алгоритм кластеризации в случае, когда не используется никакая другая вспомогательная информация. Поэтому в качестве признакового множества решено было использовать доминанты документов.

2. Тематическая близость документов и кластеризация

Пусть

$D_N = (d_1, d_2, \ldots, d_N)$ – произвольное множество документов (при этом через d_i будем обозначать как сами документы, так и их идентификаторы),

$L_i^d = (l_1^i, l_2^i, \ldots, l_j^i, \ldots, l_{n_i}^d)$ – множество доминантных лексем (доминант) i-го документа,

$\Psi_i^d = (\psi_1^i, \psi_2^i, \ldots, \psi_j^i, \ldots, \psi_{n_i}^d)$ – множество ассоциативных мощностей доминант;

$i = (1, 2, \ldots, N),\ j = (1, 2, \ldots, n_i)$.

Каждое из Ψ_i^d и L_i^d частично упорядочено по убыванию ψ так, что из $j_1 < j_2$ следует $\psi_{j_1}^k >= \psi_{j_2}^k$.

Для учета роли одинаковых доминант в различных документах перейдем от ψ_j к рангу r_j – номеру группы с одинаковыми значениями ψ_j. И в качестве «веса» (w_j^i) доминанты возьмем значение, обратное рангу:

$$w_j^i = \frac{1}{r_j^i}. \tag{1}$$

Таким образом, каждый $d_i \in D_N$ представляется векторами

$$L_i^d,\ W_i^d = (w_1^i, w_2^i, \ldots, w_j^i, \ldots, w_{n_i}^i).$$

Если $l_m^k = l_n^p$, то $w_m^k = w_n^p$ только при m=n.

Пусть $R^{k,m} = L^k \cap L^m \neq \emptyset$.

В литературе (например [14]) неоднократно отмечалось, что для решения вопроса, следует ли то или иное слово рассматривать в качестве поискового термина, необходимо учитывать контекст (cluster), в котором данное слово появляется и который, в свою очередь, может быть представлен множеством других слов.

Несмотря на «доминантность» лексем множества $R^{k,m}$, пересечение по одному слову не гарантирует тематическую близость документов. Однако именно в силу того, что из полного словаря текста отобраны доминантные лексемы, наиболее точно представляющие тему, требование $N_R > 1$, эквивалентное требованию учета контекста, может оказаться достаточным для целей определения тематической близости и автоматической кластеризации текстов.

Тогда близости документов $b_{k,m}$ и $b_{m,k}$ определяются следующим образом:

$$b_{k,m} = \sum_i^{N_R} w_i^k,\ b_{m,k} = \sum_i^{N_R} w_i^m,\ N_R > 1. \tag{2}$$

Пусть

$$B^k = (b_{k,m_1}, \ldots, b_{k,m_i} \ldots),\ m_i \neq k, \tag{3}$$

– списки близости документов $d_k \in D_N$ к другим документам множества D_N, частично упорядоченные по убыванию значений b_{k,m_i}.

Отбросив в (3) все b_{k,m_i} меньшие первых максимальных и заменив b_{k,m_i} на соответствующие идентификаторы d_{k,m_i} для каждого k, получим списки документов, максимально близких k-ым, или списки «центроидов»:

$$L^{k,max} = (d_{k,m_1}, d_{k,m_2}, \ldots, d_{k,m_i} \ldots), \tag{4}$$

$$k, m_i \in (1, 2, \ldots, N),$$

причем d_{k,m_i} в (4) обозначает идентификатор m_i - го документа, максимально близкого к k-му.

Построим начальный структурированный список кластеров, каждый элемент которого состоит из: m-го центроида, списка L_m k-ых документов с максимальными $b_{k,m}$ и числа элементов списка (списка элементов m-го кластера – C_m):

$$K = ((d_1^c, C_1, L_1), (d_2^c, C_2, L_2), \ldots (d_i^c, C_i, L_i), \ldots) \tag{5}$$

$$L_i = (d_{i,1}, d_{i,2}, \ldots).$$

Элементы списка K частично упорядочены по убыванию значения C_i.

Для получения итогового списка кластеров необходимо объединить все L_i и L_j, $i < j$, такие, что $d_j^c \in L_i$.

Таким образом, список K есть итоговое разбиения множества D_N на подмножества $(d^c \cup L_i)$, если $\forall (i < j), d_j^c \notin L_i$.

3. Эксперимент

Ниже приводятся результаты экспериментов по проверке сепарационных возможностей метода. Влияние других факторов (например размера документов) не исследовалось. Все документы взяты из Internet'a и проиндексированы системой ГИОС. Документы, не имеющие связей с другими, помещаются в кластер «разное».

В качестве контрольных, однозначно принадлежащих фиксированной предметной области, выбраны два курса лекций по СУБД. Ниже приводится содержание частей для того, чтобы дать представление читателю об их тематике.

Курс I. С.Д. Кузнецов. Введение в СУБД, 9 частей.

(Рубрикация дана точно по электронной публикации)

Часть 1. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #01/95.

1. Численные и информационные прикладные системы.
2. Файловые системы.
3. Области применения файлов.

Часть 2. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #02/95

4. Потребности информационных систем.
5. Что есть СУБД в целом – функции и структура.
6. Да, были средства (управления базами данных) в наше время...

Часть 3.СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #03/95.

Глава 4. Реляционный подход к организации баз данных, или Теория и Интуиция.

Глава 5. Базисные средства манипулирования реляционными данными, или на чем базируются языки запросов.

Часть 4. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #04/95.

Глава 6. Проектирование реляционных БД на основе принципов нормализации и семантическое моделирование баз данных.

Часть 5. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #01/96.

Глава 6. System R: более чем удачный эксперимент.

Часть 6. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #02/96.

Глава 7. Ingres: откуда пошли открытые СУБД.

Глава 8. Базы данных: и куда же все складывается?

Часть 7. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #03/96.

Глава 9. Может ли толпа людей пройти через узкую дверь и не слишком наломать бока, или Управление транзакциями в системах баз данных.

Глава 10. Надежно можно жить только имея запасы, или Журнализация изменений БД.

Часть 8. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #04/96.

Глава 11. В любое царство вводят толмачи.

Глава 12. Традиционные социальные методы в компьютерных технологиях, или СУБД в архитектуре клиент-сервер.

Глава 13. Мы не одни в этом мире, или Распределенные базы данных.

Часть 9. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #05-06/96.

Глава 14. Что день грядущий нам готовит?

Глава 15. Каждому субъекту свой объект.

Глава 16. Рулить - это от слова «действовать по правилам».

Курс II. Ладыженский. СУБД - коротко о главном.

Часть 1. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #01/95.

Введение.

Раздел 1. Реляционная база данных - основные понятия.

Часть 2. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #02/95.

Раздел 2. Сервер базы данных.

Часть 3. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #03/95.

Раздел 3. Обработка распределенных данных.

Часть 4. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #04/95.

Раздел 4. Обработка транзакций.

Раздел 5. Средства защиты данных в СУБД.

Заключение.

Литература.

Прежде всего было установлено, что, объединенные в два текста (части каждого курса собраны в соответствующий текст), они составляют один кластер.

Эти тексты, разбитые на части (всего 13), помещались в документальные среды различной тематики. Результаты кластеризации приведены ниже.

Число документов и тематическая характеристика относятся к документам среды.

3.1. Два вышеуказанных курса лекций по СУБД по частям (всего 13 текстов)

Результат. Полное разделение на два кластера, каждый из которых содержит части соответствующего курса.

Субъективная оценка «отлично».

3.2. Отдельные статьи по СУБД (всего10)

Результат. В данном случае имеет смысл полностью привести составы кластеров.

Кластер 1.

Джим Грей. Управление данными: прошлое, настоящее и будущее. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #03/98.

БД - достижения и перспективы на пороге XXI столетия. Под ред. Ави Зильбершатца, Майка Стоунбрейкера и Джеффа Ульмана. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #03/96.

С.Д. Кузнецов. Введение в СУБД. Часть2.

С. Д. Кузнецов. Введение в СУБД. Часть 1.

А.З. Ишмухаметов, В.В. Лукин. Организация словаря данных в предметно-ориентированных программных оболочках. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #01-02/98.

Э. Ларсен, Дж. Олкин, М.Портер. Oracle Media Server: предоставление потребителям интерактивного доступа к данным мультимедиа. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #01/95

Кластер 2.

К. В. Ахтырченко, В. В. Леонтьев. Распределенные объектные технологии в информационных системах. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #05-06/97.

К. В. Ахтырченко. Применение технологии Corba при построении распределенных информационных систем. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #01-02/98.

Кластер 3.

С.Д. Кузнецов. Введение в СУБД. Часть 5.

С.Д. Кузнецов. Введение в СУБД. Часть 6.

С.Д. Кузнецов. Введение в СУБД. Часть 7.

С.Д. Кузнецов. Введение в СУБД. Часть 8.

С.Д. Кузнецов. Введение в СУБД. Часть 3.

С.Д. Кузнецов. Введение в СУБД. Часть 9.

Джон М. Смит, Диана К. Смит. Абстракции баз данных: агрегация и обобщение. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #02/96.

Петер Пин-Шен Чен. Модель «сущность-связь» - шаг к единому представлению о данных. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #03/95

С.Д. Кузнецов. Введение в СУБД. Часть 4.

Кластер 4.

Ладыженский. СУБД - коротко о главном. Часть 3.

Ладыженский. СУБД - коротко о главном. Часть 2.

Ладыженский. СУБД- коротко о главном. Часть 4.

Б.А.Позин. Современные средства программной инженерии для создания открытых прикладных ИС. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #01/95

С.Д. Кузнецов. Введение в информационные системы. СИСТЕМЫ УПРАВЛЕНИЯ БАЗАМИ ДАННЫХ #02/97

Ладыженский. СУБД - коротко о главном. Часть 1.

Комментарий. Отлично. Первый кластер, по сути, введение в тему. Содержание 2-го говорит само за себя. Третий кластер – ядро, основное содержание темы СУБД. 4-й кластер, как и лекции Ладыженского в целом, имеет выраженную «ИС-доминанту».

3.3. Психология (всего текстов 25)

Результат: полное разделение на 3 кластера – психология и два кластера лекций.

Комментарий. Отлично.

3.4. Публицистика различной тематики (всего текстов 30)

Результат. 7 кластеров и «разное». Два отдельных кластера лекций.

Комментарий. Отлично.

3.5. Три различных курса лекций по философии (всего текстов 45)

Е.К. Дулуман. Философия (7). Лекции по истории натурфилософии (28). А.Н. Суворова. Введение в современную философию (10).

Результат. 6 кластеров. Два отдельных кластера лекций по СУБД.

Комментарий. Отлично.

3.6. Семь рассказов и повесть А.П.Чехова «Дама с собачкой»

Результат. 3 кластера плюс «разное». Лекции Ладыженского по-прежнему составляют отдельный кластер, *а в кластер лекций Кузнецова попадает «Дама с собачкой».*

Комментарий. Тройка. Дальнейший анализ показывает, что повесть Чехова попадает последней в список близости девятой части лекций Кузнецова из-за пересечения по доминантам «время» и «памяти».

3.7. Психология плюс «Дама с собачкой» (всего текстов 26)

Результат. Три кластера. Повесть Чехова попадает в кластер «психология». Лекции составляют два отдельных кластера.

Комментарий. Отлично. В данном случае «Дама...» попадает по месту.

3.8. Публицистика различной тематики (всего текстов 30)

Результат. Семь кластеров плюс «разное». Контрольные тексты составляют два отдельных кластера.

Комментарий. Отлично.

4. Обсуждение результатов и выводы

Представленный метод кластеризации демонстрирует высокое качество тематической сепарации текстов, что, в свою очередь, говорит о перспективности подхода, положенного в основание ассоциативной модели. По-видимому, описанный алгоритм кластеризации можно эффективно использовать для снижения доли нерелевантных документов при поиске по образцу, а также для построения субклассов после первичной классификации документов на основе заданных тезаурусов предметных областей. В основу тезаурусов могут быть положены доминантные лексемы.

Тем не менее, как показывает случай с повестью А.П. Чехова, алгоритм не гарантирует 100% тематической однородности кластеров.

Литература

1. Солтон Дж. *Динамические библиотечно-информационные системы.* М.: Мир, 1979.
2. Кузина И. *Новое поколение поисковых машин.* – http://koi.www.osp.ru/cw/1997/32/opensys/01.html
3. Керстеттер Д. *Новая лингвистическая технология повышает точность поиска* // Компьютерная неделя. N47 (121) от 2/12/1997
4. Крейнес М.Г. *Смысловой поиск и индексирование текстовой информации в электронных библиотеках: информационная технология «ключи от текста»* // Электронные библиотеки. 1999. Т 2, Выпуск 3.
5. Эссик К. *Документ - это еще не информация* // Computerworld Россия. 1998 № 25.
6. Петухов Д.А., Heuser U., Babanine A., Rosenstie W. *Применение нейронных сетей для кластеризации документов.* –http://oasis.peterlink.ru/ dap/nneng/nn−article.html
7. Нариньяни А.С. *Автоматическое понимание текста – новая перспектива.* Сайт РосНИИ искусственного интеллекта: http://www.rriai.org.ru
8. Кузин Л.Т. *Основы кибернетики.* М.: Энергия, 1979.
9. *Искусственный интеллект.* - В 3-х кн. Кн.2. Модели и методы: Справочник / Под ред. Д.А. Поспелова. М.: Радио и связь, 1990.
10. Загоруйко Н.Г. *Прикладные методы анализа данных и знаний.* Новосибирск: Изд-во Ин-та математики, 1999.
11. Бонгард М.М. *Проблема узнавания.* М.: Физматгиз, 1967.
12. Чанышев О.Г. *Ассоциативная модель естественноязыкового текста* // Вестник Омского университета. 1977. Вып. 4. С.17-20.

13. Чанышев О.Г. *Ассоциативная модель реального текста и ее применение в процессах автоиндексирования* // Труды Седьмой национальной конференции по искусственному интеллекту с международным участием КИИ'2000. - Москва: Изд-во Физико-математической литературы, 2000. С. 430-438.
14. Bookstein A.,S Klein . T. *Clumping Properties of Content-Bearing Words* // JASIS. 1998. №2.