КЛАСТЕРИЗАЦИЯ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ МЕТАИНФОРМАЦИИ

DOCUMENT CLUSTERING USING METADATA

С.Г. Баглей *(baglei@galaktika.ru)*,
А.В. Антонов *(alexa@galaktika.ru)*,
В.С. Мешков *(meshkov@galaktika.ru)*,
А.В. Суханов *(sukhanov@galaktika.ru)*

Корпорация "Галактика", Москва

В статье описывается подход к кластеризации документов, реализованный в поисково-аналитической системе Галактика-Зум на базе модифицированного алгоритма LSA. Основная задача, которая решается с помощью описываемого подхода – разделение множества документов на области-кластеры по общности тем, то есть, по сходству векторов признаков. В отличие от традиционной реализации алгоритма LSA, базовыми единицами для проведения кластеризации являются слова и словосочетания, составляющие ИнфоПортрет документов. Элементами ИнфоПортрета являются языковые инварианты, статистически отличающие данную выборку документов.

Введение

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

Основным понятием в системе Галактика-Зум 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 отдельных 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, то есть, может рассматриваться как *метаинформация*, соответствующая документу. Принимая данное условие, перейдем к описанию проблемной области, рассматриваемой в статье.

Общей проблемой, снижающей эффективность работы пользователя с поисковой системой, является избыточность информации при выдаче результатов по запросу. Причинами возникновения избыточности могут быть, например, нечетко сформулированные запросы к поисковой системе, омонимичность элементов поискового предписания и другие.

Задача уменьшения избыточности, может решаться различными способами. Достаточно эффективным среди них является диалог пользователя с системой, то есть, режим, при котором пользователю предоставляется возможность уточнения своих информационных предпочтений. Кластеризация выборки документов представляет собой эффективное средство повышения качества диалога, позволяющее проводить разбиение полученной выборки по тематическим признакам. Далее рассматривается метод кластеризации, реализованный в системе Галактика-Зум.

В качестве основы метода выбран известный алгоритм кластеризации LSA/LSI, использующий принципы факторного анализа для выявления латентной структуры объектов. Задачей факторного анализа является выделение главных факторов из пространства элементарных. Выбор данного алгоритма обусловлен рядом причин. Во-первых, LSA не нуждается в обучении. То есть, при кластеризации формируется такая структура кластеров, которая зависит исключительно от обрабатываемых данных. Кроме того, не требуется проведения этапа предварительной настройки алгоритма. Во-вторых, из опыта предыдущих работ [5], метод LSA признается лучшим для выявления латентных зависимостей в структуре объектов.

Обозначим предметную область работы традиционного алгоритма кластеризации LSA. Основой работы служат объекты, представляющие собой слова или термины документа. То есть, те слова, из которых состоит документ, являются элементарными признаками для проведения кластеризации, множество которых составляет пространство признаков. Каждый документ из множества документов, предназначенных для кластеризации,

является вектором в пространстве признаков. В качестве недостатка такого подхода можно признать принимаемое в LSA допущение, что все термины в документе имеют одинаковую значимость. Дальнейшие вычисления в пространстве признаков производятся исходя из этого упрощения, что негативно сказывается на результатах работы алгоритма.

Система Галактика-Зум позволяет не прибегать к условию равнозначности слов. Преимуществами, предоставляемыми системой с точки зрения задачи кластеризации являются следующие:

- 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.

Используя данные преимущества, перейдем к формальному описанию задачи кластеризации и ее решению.

Постановка задачи

Используется следующая модель задачи кластеризации.

Ω - множество документов (объектов распознавания) – пространство образов.

$\omega \in \Omega$ - документ (образ).

$f(\omega) : \Omega \to M, M = \{1,2,...,m\}$ - неизвестная наблюдателю индикаторная функция, разбивающая множество документов Ω на m непересекающихся подмножеств (кластеров), параметры которых заранее неизвестны $\Omega^1, \Omega^2, ..., \Omega^m$. Количество кластеров может быть произвольным или фиксированным. Условие о непересекаемости подмножеств может не выполняться в условиях частичного совпадения наборов параметров, относящихся к разным подмножествам.

X - пространство измерений, воспринимаемых наблюдателем (пространство признаков).

$x(\omega) : \Omega \to X$ - функция, ставящая в соответствие каждому документу ω точку $x(\omega)$ в пространстве признаков. Вектор $x(\omega)$ - это образ документа, воспринимаемый наблюдателем. В пространстве признаков существуют, заранее неизвестные, непересекающиеся множества точек $K_i \subset X, i = 1,2,...,m$, соответствующих документам одного кластера.

$\hat{f}(x) : X \to M$ - решающее правило – оценка для $f(\omega)$ на основании $x(\omega)$, т.е. $\hat{f}(x) = \hat{f}(x(\omega))$.

Пусть $x_j = x(\omega_j), j = 1,2,...,N$ - доступная наблюдателю информация о функциях $f(\omega)$ и $x(\omega)$, но сами эти функции наблюдателю неизвестны.

Задача заключается в построении такого решающего правила $\hat{f}(x)$, чтобы разделение документов на кластеры по сходству соответствующих векторов признаков проводилось с минимальным числом ошибок - минимизация потерь от неправильного распознавания.

Решение

Пусть задана выборка N документов, в которой каждый документ представляется последовательностью словоформ. Выборке соответствует ИнфоПортрет – множество значимых слов, которое составляет *пространство признаков* X. *Множество документов* – это множество точек или векторов этого пространства. Координатами точки $x_\bullet$ являются величины значимости каждого элемента ИнфоПортрета для данного документа: вклад признака в близость ИнфоПортретов. Величина значимости задается следующей формулой:

$x_\bullet = M_\bullet \times D_\bullet \times f_\bullet$, где:

$M_\bullet$ - основная составляющая вклада признака в близость ИнфоПортретов,

$D_\bullet$ - невязка близости ИнфоПортретов,

$f_\bullet$ - фильтрующий множитель.

Каждому документу ставится в соответствие единственное значение вектора признаков и наоборот: каждому значению вектора признаков соответствует единственный документ. Координаты документов в пространстве признаков образуют матрицу A.

В качестве решающего правила предлагается использовать метод обнаружения латентных связей LSA/LSI, который является реализацией основных принципов факторного анализа применительно к множеству документов.

Матрица A может быть разложена на произведение трех матриц (сингулярное разложение) следующим образом:

$A = S\Lambda D^{'}, где$

A - исходная матрица размера $I \times N$, I - количество признаков в ИнфоПортрете,

Λ - диагональная матрица размера $m \times m$ ($m \le r = rank(A)$) и содержащая собственные значения матрицы A,

S, D - матрицы левых (признаки) и правых (документы) собственных векторов матрицы A, соответственно размера $I \times m$ и $N \times m$, т.ч. $S^{'}S = D^{'}D = I_m$.

Документы и признаки, проецируясь на m-мерное факторное пространство посредством матриц D и S, образуют области - кластеры.

Результаты экспериментов

Для оценки качества работы метода нами был проведен ряд экспериментов. Далее приведены результаты одного из них. В качестве основы для его проведения использовался массив документов, состоящий из газетных и журнальных статей в базе системы Галактика-Зум. В ходе эксперимента моделировалась ситуация проведения реального поискового запроса.

Был проведен следующий запрос: Кисин или (космос и (катастрофы или аварии) и (космонавт или астронавт) и ctx(открытый космос)) или (проститутка и бордель).

В табл. 1 приведены характеристики базы и полученной выборки:

Параметр	Количество, млн. ед.
Количество документов в базе	1,7
Количество слов в базе	6,4
Число словомест в базе	882
Число словомест в выборке	0,4
Количество словосочетаний в базе	3,8
Число мест словосочетаний в базе	83
Число мест словосочетаний в выборке	0,05

Таблица 1. Численные характеристики базы и выборки

В результате проведенного запроса была получена выборка и сформирован ее ИнфоПортрет, верхними элементами которого были слова и словосочетания, приведенные в табл. 2.

ПРОСТИТУЦИЯ
ШАТТЛ
ОТКРЫТЫЙ КОСМОС
МКС
КОСМИЧЕСКИЙ
КОСМОС
КИСИН
ДИСКАВЕРИ
ПУБЛИЧНЫЙ ДОМ
СУТЕНЕР
КОРАБЛЬ
АСТРОНАВТ
ПРИТОН
ПОЛЕТ
ОРБИТА
КОСМИЧЕСКАЯ СТАНЦИЯ
КОСМОНАВТ
ЭКИПАЖ
ЧЕЛНОК
СЕКСУАЛЬНЫЕ УСЛУГИ
КОСМИЧЕСКИЙ КОРАБЛЬ
КРАСНЫЕ ФОНАРИ
ИНТИМНЫЕ УСЛУГИ
ТОПЛИВНЫЙ БАК
БАЛЬЗАМ
СКАФАНДР
ЖИВОЙ ТОВАР

ЖРИЦА
НАСА
НАШ КОММЕНТАРИЙ
СЕКСУАЛЬНЫЙ
ПУТАНЫЙ
КОЛУМБИЯ
ТОРГПРЕДСТВО
СЕКС
ЛЕГКОЕ ПОВЕДЕНИЕ
КОСМИЧЕСКИЙ ЧЕЛНОК
ПИЛОТИРОВАТЬ
ПОДПОЛЬНЫЙ БОРДЕЛЬ
МОИ БАЛЬЗАМЫ
ОБШИВКА
КОСМИЧЕСКОЕ АГЕНТСТВО
ПЛОТНЫЕ СЛОИ
ПОЛИЦИЯ
АМЕРИКАНСКИЕ ШАТТЛЫ
КОСМОНАВТИКА
БАЙКОНУР
ПРОДАЖНАЯ ЛЮБОВЬ
КОСМОДРОМ

Таблица 2. Верхние элементы ИнфоПортрета выборки

Рассматривая работу алгоритма кластеризации как часть функциональности поисковой системы Галактика-Зум, в качестве исходных рубрик были определены документы полученной выборки, отвечающие следующим условиям:

- ранг документа в выборке не должен быть меньше выбранного порога, принятого, в нашем случае, размером в 150 документов;
- документ из выборки должен быть отнесен в результате экспертной оценки к одному из элементов запроса:

1)*Кисин;*

2)*космос и (катастрофы или аварии) и (космонавт или астронавт) и ctx(открытый космос)*

3)*проститутка и бордель.*

Далее была проведена кластеризация документов, полученных по запросу. Для оценки эффективности предложенного метода мы оценили как результаты его работы, так и результаты кластеризации полученной выборки с использованием традиционного алгоритма LSA/LSI [6]. В качестве модели документа в традиционном методе мы также использовали ИнфоПортрет, формируемый в системе Галактика-Зум.

	Система Галактика-Зум	LSA/LSI
Количество кластеров	3	4
Количество документов в выборке	308	500
Количество документов, включенных в кластеры	112	65
Количество документов, общих для кластеров	0	41
Минимальное число документов в кластере	12	6
Максимальное число документов в кластере	69	47
Минимальное число объектов в кластере	6	20
Максимальное число объектов в кластере	32	39
Среднее отклонение числа документов в кластере	37	27
Среднеквадратичное отклонение	29	17

числа документов в кластере		
Среднее отклонение числа объектов в кластере	22	32
Среднеквадратичное отклонение числа объектов в кластере	14	8.3
Коэффициент вариации документов	0.78	0.63
Коэффициент вариации объектов	0.64	0.26

Таблица 3.Основные параметры кластеризации выборки по запросу

После проведения первого этапа кластеризации в массиве были выделены кластеры документов с соответствующими ИнфоПортретами. Далее приводятся ИнфоПортреты полученных кластеров, упорядоченных по близости ко всей выборке документов.

Кластер №1	
Галактика-Зум	LSA/LSI
Количество документов в кластере: 69	Количество документов в кластере: 47
ПОЛЕТ	НЯНЯ
КОСМОС	ВОВЛЕЧЕНИЕ
МКС	НЕДЕЛЬНЫЙ
ЧЕЛНОК	УБОП
СКАФАНДР	АСТРОНАВТ
ШАТТЛ	ОХРАННИК
ДИСКАВЕРИ	ПОВРЕЖДЕНИЕ
КОСМИЧЕСКИЙ	МУЖИК
БАЙКОНУР	СИФИЛИС
КОРАБЛЬ	КЕННЕДИ
КОСМОНАВТ	ХОЛОКОСТ
КОСМОДРОМ	ПОЛИЦЕЙСКИЙ
ЭКИПАЖ	ФОНАРЬ
АСТРОНАВТ	ФРАГМЕНТ
ГИРОСКОП	НОЧНОЕ
ПОЛОТЬ	НРАВСТВЕННОСТЬ
ОТКРЫТЫЙ КОСМОС	ВИДЕОКАМЕРА
ГАГАРИН	РЕЖИССЕР
НОГУТИ	НАГРЯНУТЬ
ПИЛОТИРОВАТЬ	КОЛЛИНЗ
НАСА	КАРАТЬ
ЛЮК	ИНЦИДЕНТ
РОБИНСОН	ГАГАРИН
ЛЕОНОВ	БЕЗВРЕДНЫЙ
ОРБИТА	ВИДЕОСЪЕМКА
ШЛЮЗ	ШОСТАКОВИЧ
МАРС	СПИВАКОВ
КРИКАЛЕВ	ТАНЦЕВАТЬ
СТЫКОВОЧНЫЙ ОТСЕК	СУТЕНЕР
ЭКСПЕДИЦИЯ	САУНА
КОСМОНАВТИКА	ДИАНА
КОЛУМБИЯ	НОВОГОДНИЙ
	ТЕРМОИЗОЛЯЦИОННЫЙ
	ВРАЩЕНИЕ
	АПОЛЛОН
	АНРИ

Таблица 4. Элементы ИнфоПортретов кластеров №1 в системе Галактика-Зум и традиционном LSA

Кластер №2	
Система Галактика-Зум	LSA/LSI
Количество документов в	Количество документов в кластере:

кластере: 31	27
ПРОСТИТУЦИЯ	ЧЕЛНОКОВ
ПОЛИЦИЯ	АСТРОНАВТ
ТОРГПРЕДСТВО	ИЛЛЮМИНАТОР
СУТЕНЕР	ПОЛЕТ
ПРИТОН	ШАТТЛ
НОЧНОЙ	ВЗЛЕТ
	ДИСКАВЕРИ
	ТОПЛИВНЫЙ
	ЭКИПАЖ
	АТЛАНТИС
	ОБШИВКА
	БАК
	ПОЛОТЬ
	КОРАБЛЬ
	ПОСАДКА
	ПОВРЕЖДЕНИЕ
	СТЫКОВКА
	БОРТ
	КОЛУМБИЯ
	МОДУЛЬ
	КОСМИЧЕСКИЙ
	ЗАПУСК
	ФРАГМЕНТ
	ЦУП
	ДНИЩЕ
	ЧЕЛНОК
	ПРИЗЕМЛИТЬСЯ
	МКС
	НАСА
	ПОВРЕДИТЬ
	КОЛЛИНЗ
	КЕННЕДИ
	ГРУЗОВОЙ
	ОТОРВАТЬСЯ
	РОБИНСОН
	ВВС
	МЫС
	КЕРАМИЧЕСКИЙ
	ОТВАЛИТЬСЯ

Таблица 5. Элементы ИнфоПортретов кластеров №2 в системе Галактика-Зум и традиционном LSA

Кластер №3	
Система Галактика-Зум	LSA/LSI
Количество документов в кластере: 27	Количество документов в кластере: 6
ОРКЕСТР	НЕДЕЛЬНЫЙ
КИСИН	БЕЗВРЕДНЫЙ
ОПЕРА	БАРСУЧИЙ
СКРИПКА	АНАЛЬГЕТИК
БЕТХОВЕН	ПОЯСНИЦА
ДИРИЖЕР	ПРОТИВО ВОСПАЛИТЕЛЬНЫЙ
ПИАНИСТ	БАЛЬЗАМ
СОРОКИН	ОБЕЗБОЛИВАЮЩИЙ
КОМПОЗИТОР	ОБОСТРЕНИЕ
КАМЕРНЫЙ ОРКЕСТР	РАДИКУЛИН
КОНСЕРВАТОРИЯ	АРТРИТ
ПРОКОФЬЕВА	КРОВОТОК

БОЛЬШОЙ ЗАЛ	ДИКУЛЯ
ПРОКОФЬЕВ	МЕДВЕЖИЙ
БОЛЬШОЙ ТЕАТР	СУСТАВ
КАМЕРНЫЙ	ЭКСТРАКТ
ВЕНГЕРОВА	ДИКУЛЬ
КОСМОС	ПОЗВОНОЧНИК

Таблица 6. Элементы ИнфоПортретов кластеров №3 в системе Галактика-Зум и традиционном LSA

Кластер №4 Количество документов в кластере: 26 LSA/LSI
ПОЛЕТ
КОСМИЧЕСКИЙ
ОКОЛОЗЕМНЫЙ
ОТРАБОТАТЬ
ПОЛОТЬ
КОСМОС
МОДУЛЬ
ЛУНА
ГАГАРИН
ГАГАРИНА
ЭКИПАЖ
МЕДИКО-БИОЛОГИЧЕСКИЙ
КОРАБЛЬ
ВОВЛЕЧЕНИЕ
ОРБИТАЛЬНЫЙ
ЗВЕЗДНЫЙ
ЗАРАБОТОК
ВЕРБОВКА
ЗЕМНОЙ
РЕЖИССЕР
НЕСОВЕРШЕННОЛЕТНЯЯ
ПИЛОТИРОВАТЬ
СУТЕНЕР
НЕВЕСОМОСТЬ
ШОСТАКОВИЧ
КОСМОНАВТИКА
ПРОКОФЬЕВА
АВТОМАТИЧЕСКИЙ
КОСМОНАВТ
ИЗБИВАТЬ

Таблица 7. Элементы ИнфоПортретов кластера №4 в традиционном LSA/LSI

При сравнении результатов кластеризации необходимо учитывать, что в обоих случаях: как при использовании предлагаемого подхода, так и традиционного LSA/LSI применялась модель документа, формируемая в системе Галактика-Зум и характерная для нее. С учетом данного обстоятельства можно предположить, что алгоритм LSA/LSI показал несколько худшие результаты, чем при использовании модели документа, изначально принятой для метода. Использование ИнфоПортрета как метода фильтрации объектов при проведении факторного анализа, а также словосочетаний наряду с терминами, как это принято в LSA/LSI, является существенным преимуществом нашего подхода. Подобное представление модели документа вполне себя оправдало полученными результатами кластеризации.

При экспертном анализе документов, входящих в кластеры, было оценено соответствие полученных документов условным рубрикам, указанным выше. Результаты оценки приведены в таблице 8.

	Общее количество документов		Количество документов, относящихся к рубрике	
№ кластера	ГЗ	LSA/LSI	ГЗ	LSA/LSI
1	69	47	69	26
2	31	27	31	27

3	27	6	27	0
4		26		19

Таблица 8. Результаты экспертной оценки документов в кластерах

При кластеризации с помощью метода, принятого в системе Галактика-Зум, каждый из сформированных кластеров был отнесен к различным рубрикам. То есть, структура рубрик была полностью воспроизведена.

Заключение

Мы применили алгоритм LSA для кластеризации документов в системе Галактика-Зум, используя возможности получения метаинформации документов, предоставляемые системой. Полученные результаты кластеризации мы сравнили с традиционным подходом LSA/LSI. Исходя из результатов сравнения, можно сделать вывод, что предлагаемый нами метод показывает лучшие результаты по сравнению с традиционным LSA/LSI. Задача разбиения документов и объектов ИнфоПортрета на кластеры успешно решается. Качество кластеризации существенно возросло по сравнению с традиционной реализацией алгоритма LSA/LSI, примененного ранее для решения данной задачи. Таким образом, использование модифицированного метода себя оправдало.

Вместе с тем, в качестве перспективных задач по улучшению качества работы алгоритма можно отметить необходимость повышения полноты кластеризации, и, возможно, расширение ИнфоПортрета кластеров.

Список литературы

1. *Антонов А.В. Методы классификации и технология Галактика-Zoom // сб. 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 №8, 2001.*
4. *Антонов А., Мешков В. Современные проблемы поисковых систем и некоторые пути их преодоления // Сер. «Аналитика-Капитал», Москва, 2000.*
5. *Кириченко К.М, Герасимов М.Б. Обзор методов кластеризации текстовых документов // Материалы международной конференции Диалог'2001.*
6. *Deerwester, S., Dumais, S., Landauer, T., Furnas, G. and Harshman, R. Indexing by latent semantic analysis // Journal of the Society for Information Science, 1990, vol. 41(6), 391-407*