Сравнительный анализ результатов кластеризации на основе однословных и составных ключевых термов

Д.А. Ткачев
Институт вычислительных технологий СО РАН
Новосибирск, Россия

Постановка задачи

1. **Выполнение процесса кластеризации электронных документов**

   - на основе многомерного шкалирования, в частности с использованием данных о ключевых словосочетаниях

   - **выделение ключевых словосочетаний** из корпусов текстов произвольной тематики, без знания тезауруса предметной области

2. **Оценка качества кластеризации**

   - Вычисление внешних мер

   - “ручная” проверка случайных кластеров

Координатное индексирование

Преимущества автоматизации индексирования:

- обеспечивает единообразие индексирования, почти невозможное для человеческого интеллекта;
- обходится существенно дешевле/быстрее
- демонстрирует хорошие результаты

Индексирование на базе тезауруса:

плюсы:

- качество выполнения

минусы:

- необходимость составления тезауруса и словаря словоформ
- большие затраты машинного времени

Морфологический анализ

Морфологический анализ документа – Mystem

Яндекс, бесплатна для некоммерческого использования

http://company.yandex.ru/technology/mystem/

Выделение ключевых слов

фильтрация на стоп слова

фильтрафия на имена собственные

Выделение ключевых фраз по шаблонам:

[Причастие] [Существительное]

[Прилагательное] [Существительное]

[Существительное] [Существительное в творительном падеже]

[Существительное] [Существительное в родительном падеже]

Отсечение отобранного количества ключевых термов

Ограничение множества

Условие ограничения множества ключевых термов/фраз

$$\textbf{KeyPhrase } i: \quad \frac{MAX(Frequency)}{Frequency(i)} < \frac{word\_num}{3}$$

Где:

MAX(Frequency) – максимальная частота встречаемости 1 из множества термов

Frequency(i) – частота встречаемочти i-го терма

word_num – желаемое (ориентировочное) количество отобранных термов

Примеры выделенных термов и фраз

Таблица 1. Выделенные термы из романа Л.Н. Толстого «Война и мир»

князь -	2011	княжна марья -	93 (княжною Марьей)
человек -	1755	старый князь -	92 (старого князя)
княжна -	885	полковой командир -	76 (полкового командира)
граф -	734	старый граф -	53 (старого графа)
время -	714	русский армия -	50 (русская армия)
москва -	644	русский войска -	41 (русскими войсками)
француз -	595	молодой человек -	32 (молодого человека)
государь -	591	исторический лицо -	30 (исторические лица)
солдат -	581	выражение лицо -	30 (выражением лица)
наполеон -	575	французский армия -	28 (французской армией)
жизнь -	572	главный квартира -	27 (главная квартира)
слово -	566	французский войска -	26 (французские войска)
рост -	544	старый графиня -	23 (старой графини)
офицер -	543	князь андрей -	23 (князем Андреем)
кутузов -	533	военный министр -	23 (военного министра)
армия -	463	французский офицер -	21 (французских офицеров)
лошадь -	450	великий князь -	20 (великого князя)
графиня -	441	расположение дух -	19 (расположении духа)
войска -	435	лицо наташа -	19 (лицо Наташи)

Метод кластеризации

Использовался “жадный” алгоритм:

- 1 построение матрицы сходства каждого документа с каждым (*N*x*N*)
- 2 выбор строки с максимальной суммой элементов (выбор центроида)
- 3 определение документов входящих в данный кластер
- 4 исключение из матрицы всех строк и столбцов по выбранным документам
- повторение шагов 2-4 пока есть не кластеризованные документы

Вычисление меры сходства между документами:

$$P(d_1,d_2)=\frac{\sum_{i\in K(d_1)\cap K(d_2)}(n(d_1,t)+n(d_2,t))}{n(d_1)+n(d_2)} \qquad (1)$$

P(*d1*,*d2*) в диапазоне [0, .. , 1]

Вычислительные эксперименты

Серии экспериментов:

1.Кластеризация множества документов правовой направленности, около 1300 документов

2.Набор научных документов математической направленности, содержащих классификационные признаки классификатора MSC2000, около 300 документов

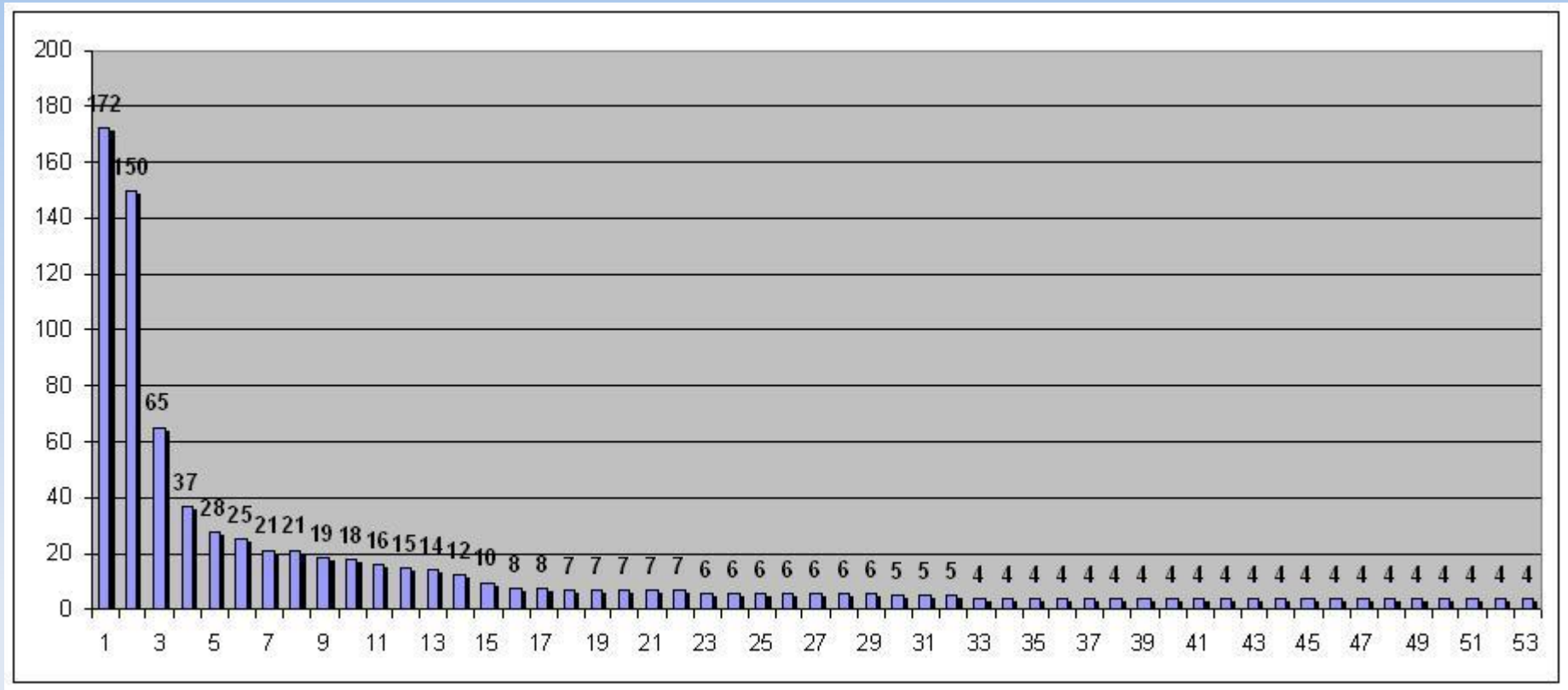
Примеры полученных кластеров

Кластер, общая тематика – налогооблажение и уклонение от уплаты
1. Бухгалтер в России.
2. Функции государства - налогообложение и взимание налогов.
3. Налоговые преступления.
4. Уклонение от уплаты налогов с организаций.
5. Уклонение физического лица от уплаты налога или страхового взноса.

Кластер, общая тематика – управление и госслужба
1. Понятие, принципы и порядок прохождения государственной службы.
2. Управление: основные понятия, система управления, ее признаки [...].
3. Основные принципы создания, [...] организации арбитражных управляющих.
4. Особенности государственной службы субъекта Российской Федерации.
5. Органы внутренних дел Российской Федерации, правовые основы [...].

Кластер, общая тематика – имущественные права
1. Институциональные аспекты землепользования.
2. О возможности защиты права собственности на недвижимость путем виндикации.
3. Природа виндикационного притязания и элементы виндикационного иска.
4. Правовое положение лица, владеющего имуществом [...].
5. Критика понятия «объект правоотношения».

Распределение документов



Документов, не включенных ни в какой кластер, оказалось **202**, т. е. около **15%** от общего количества документов

Оценивание мер качества

Оценка мер качества кластеризации:

1.Кластеризация подготовленных текстов жадным алгоритмом с различными значениями порогового параметра

2.Получение результатов кластеризации с использованием однословных ключевых терминов и результатов, основанных на смешанном критерии

3.Вычисление внешних мер для полученных результатов

Вычисление мер

Таблица 2. Коэффициенты для подсчета внешних мер сходства

Для каждой пары документов d_j и d_i	d_j и d_i принадлежат одному кластеру в «эталонном» разбиении	d_j и d_i принадлежат разным кластерам в «эталонном» разбиении
d_j и d_i принадлежат одному кластеру в автоматическом разбиении	a	c
d_j и d_i принадлежат разным кластерам в автоматическом разбиении	b	d

Вычисление мер (продолжение)

$$\mathrm{Re}call = \frac{a}{a+b}$$

$$Precision = \frac{a+d}{a+b+c+d}$$

$$Error = \frac{b+c}{a+b+c+d}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Результаты экспериментов

Таблица 3. Меры, на основе анализа одиночных ключевых слов

Анализ на основе кодов классификатора 2-го уровня (вида 76Mxx)			
Мера	Значение параметра (величина схожести)		
	0.4	0.5	0.6
Recall	0.2263	0.2394	0.3223
Precision	0.7749	0.8201	0.8464
Error	0.2251	0.1799	0.1536
F_1	0.3503	0.3707	0.4668

Таблица 4. Меры на основе анализа одиночных ключевых слов и двухсловных выражений

Анализ на основе кодов классификатора 2-го уровня (вида 76Mxx)			
Мера	Значение параметра (величина схожести)		
	0.4	0.5	0.6
Recall	0.2771	0.3796	0.4624
Precision	0.8268	0.8493	0.8527
Error	0.1732	0.1507	0.1473
F_1	0.4150	0.5247	0.5996

Спасибо за Ваше внимание!

Д.А. Ткачев
Институт вычислительных технологий СО РАН
Новосибирск, Россия