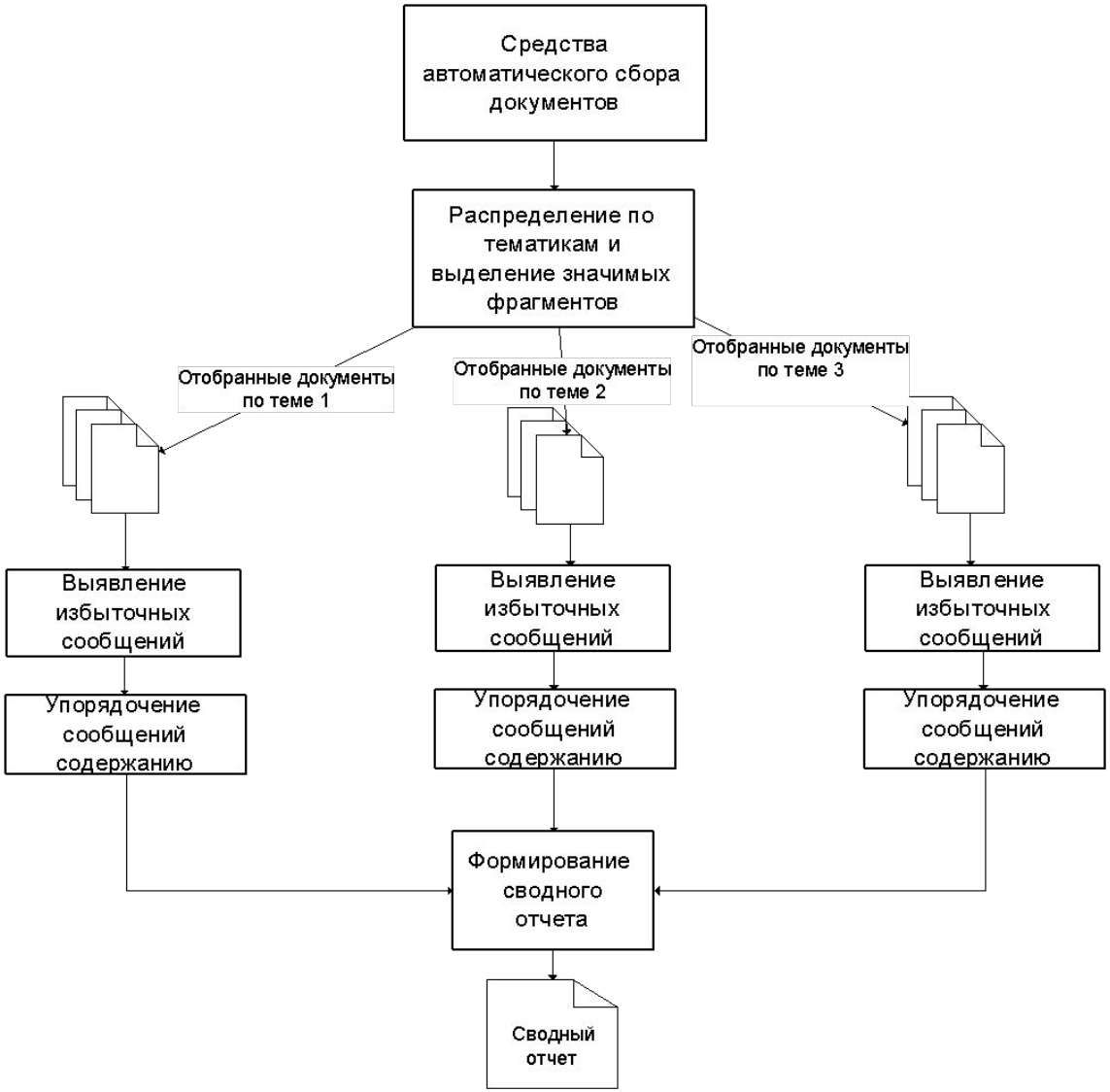
Тематическое упорядочение текстов при формировании сводных документов

Васильев В.Г.
ООО «ЛАН_ПРОЕКТ»
vvg_2000@mail.ru

Технология формирования сводных документов



Пример тематического упорядочения

ФРАГМЕНТ НЕУПОРЯДОЧЕННОГО СПИСКА	ФРАГМЕНТ УПОРЯДОЧЕННОГО СПИСКА
Более 3,5 тыс. задействованных на вредных производствах жителей Новороссийска ….	***На реконструкцию рентгенкабинетов Псковской области выделено 6,7 млн рублей***
В Магнитогорске формируется медико-социальная программа геронтологической …	***На реконструкцию рентгенкабинетов Псковской области выделено 6,7 млн рублей***
В Новосибирске будет разработана стратегия развития системы здравоохранения на …	***6,7 млн. рублей выделено из областного бюджета на рентген кабинеты***
Диагностическое оборудование, поступающее в Омскую область в рамках нацпроекта …	**Диагностическое оборудование, поступающее в Омскую область в рамках нацпроекта …**
Бороться с раком помогает национальный проект Здоровье	***Получено 149 единиц диагностического оборудования***
Костромская область в 2007 г. в рамках нацпроекта Здоровье получит 43 машины	***С начала года 31 житель Новороссийска получил высокотехнологичную медпомощь***

Задача тематического упорядочения текстов

Постановка задачи

Требуется найти такую перестановку $(l_1, \ldots, l_n)$ текстов $\{1, \ldots, n\}$, что тематически близкие тексты будут находиться рядом друг с другом, а тематически отличные – далеко.

Особенности

- Небольшое число текстов (100 – 1 000);
- Ограниченное время;

Методы решения

- Методы решения задачи коммивояжера;
- Методы решения задачи размещения элементов;
- Методы иерархического кластерного анализа;
- Методы ранжирования текстов.

Задача коммивояжера

Требуется найти такую перестановку $(l_1, \ldots, l_n)$ номеров текстов $\{1, \ldots, n\}$, на которой достигается максимум функции

$$C_{TSP}(l_1, \ldots, l_n) = \sum_{i=2}^{n} sim(x_{l_{i-1}}, x_{l_i}),$$

где $sim(x, y)$ - мера близости между текстами x и y.

Приближенные алгоритмы:

- жадные – алгоритм ближайшего соседа (NN);
- локальной оптимизации – метод двойного выбора (2OPT);
- случайного поиска – генетический алгоритм (GA).

Вычислительная сложность не более $O(n^3)$

Методы решения задачи одномерного размещения элементов

Требуется найти такую перестановку $(l_1, \ldots, l_n)$ номеров текстов $\{1, \ldots, n\}$, на которой достигается минимум функции

$$C_{PPP}(l_1, \ldots, l_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} |j - i| sim(x_{l_i}, x_{l_j})$$

Приближенные алгоритмы:
- локальной оптимизации – метод перестановки соседних элементов (PPP).

Вычислительная сложность - $O(n^3)$

Метод на основе иерархического кластерного анализа

Схема алгоритма

1. Построить бинарное дерево вложенных классов путем восходящего иерархического кластерного анализа.
2. Выполнить вращение элементов классов снизу-вверх таким образом, чтобы близкие элементы в соседних классах были бы крайними элементами.
3. Сформировать перестановку текстов, путем выполнения обхода переупорядоченного дерева сверху вниз.

Вычислительная сложность - $O\left(n^2\right)$

Метод на основе спектрального анализа (типа PageRank)

Пусть W - матрица близости текстов $n \times n$,

$$D = diag(d_1,...,d_n),\ d_i = \sum_{j=1}^{n} w_{ij},\ S = D^{-1/2} W D^{-1/2}.$$

Вектор итогового ранжирования текстов $f^* = (f_1,...,f_n)$ находится итерационной процедурой:

$$f(t+1) = \alpha S f(t) + (1-\alpha) y,\ t = 1,...,t_{\max}.$$

Вычислительная сложность - $O(n^3)$

Модели представления и вычисления близости текстов

Обозначение	Описание модели
TFIDF	Стандартная теоретико-множественная модель Косинусная мера близости
KGRAMM	Текст представляется хэш-кодами всех последовательностей слов длины k. Мера включения
KGRAMM TFIDF	Комбинированная модель и мера близости

Тестовые массивы

Название массива	Число текстов	Число рубрик	Комментарий
Yandex News	256	21	Одноуровневый набор сюжетов
Google News	511	24	Двухуровневый набор сюжетов
Reuters 21578-6	935	6	Подмножество из 6 рубрик массива Reuters-21578
ROMIP 2004	2000	173	Нормативно-правовые документы РОМИП 2004
Reuters 21578	5000	142	Массив Reuters-21578

Показатели качества тематического упорядочения

Суммарная дисперсия

$$C_{\text{var}}(l_1,\ldots,l_n)=\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i\in\omega_j}\left(l_i-\bar{l}_j\right)^2\right)$$

Нормированная суммарная дисперсия

$$C(l_1,\ldots,l_n)=\frac{\sum_{j=1}^{r}\left(n_j^2-1\right)}{12\sum_{j=1}^{r}\left(\frac{1}{n_j}\sum_{i\in\omega_j}\left(l_i-\bar{l}_j\right)^2\right)}$$

где $\omega_1,\ldots,\omega_k$ - эталонные классы, $(l_1,\ldots,l_n)$ - итоговая перестановка текстов, n_j - число документов в класса ω_j.

Время работы алгоритмов тематического упорядочения

Время (сек)	HIER	2OPT	Rank	NN	GA	PP
Yandex News	0.06	0.08	0.02	0.03	19	2.7
Google News	0.1	0.65	0.1	0.09	34	3.5
Reuters 21578_6	0.22	4.72	0.36	0.24	64	8.67
Romip 2004	1.7	86	2.5	1.8	133	90
Reuters 21578	13	>1500	36	6	398	>1500

Качество тематического упорядочения

Алгоритм	Модель текста	Yandex News	Google News	Reuters 21578_6	Romip
PPP	TFIDF	0.0176	0.0148	0.0375	0.0014
	KGRAMM TFIDF	0.0246	0.0186	0.0375	0.0014
2OPT	TFIDF	0.0184	0.0163	0.0377	0.0015
	KGRAMM	0.0034	0.0104	0.0359	0.0014
	KGRAMM TFIDF	0.0254	0.0189	0.0377	0.0015
2OPT TAIL	TFIDF	**0.0350**	**0.0979**	0.0387	0.0015
	KGRAMM	0.0036	0.0088	0.0372	0.0014
HIER	KGRAMM	0.0048	0.0148	0.0363	0.0015
	TFIDF	0.0160	0.0519	**0.0427**	0.0015
	KGRAMM TFIDF	0.0158	0.0492	0.0427	0.0015
NN	KGRAMM	0.0037	0.0104	0.0348	0.0016
	TFIDF	0.0072	0.0506	0.0392	0.0013
	KGRAMM TFIDF	0.0081	0.0515	0.0406	0.0015
RANK	TFIDF	0.0049	0.0372	0.0417	0.0016
GA	TFIDF	0.0087	0.0089	0.0347	0.0014

Заключение

Перспективные направления исследований :

- учет специфики документов различных типов (новости, электронная почта, служебных документы, научные работы);
- использование других методов для тематического упорядочения (проецирования, нейронные сети);
- устранения повторяющихся фрагментов в различных текстах из заданного массива.