Автоматическое формирование рубрикатора полнотекстовых документов

Пескова Ольга Вадимовна

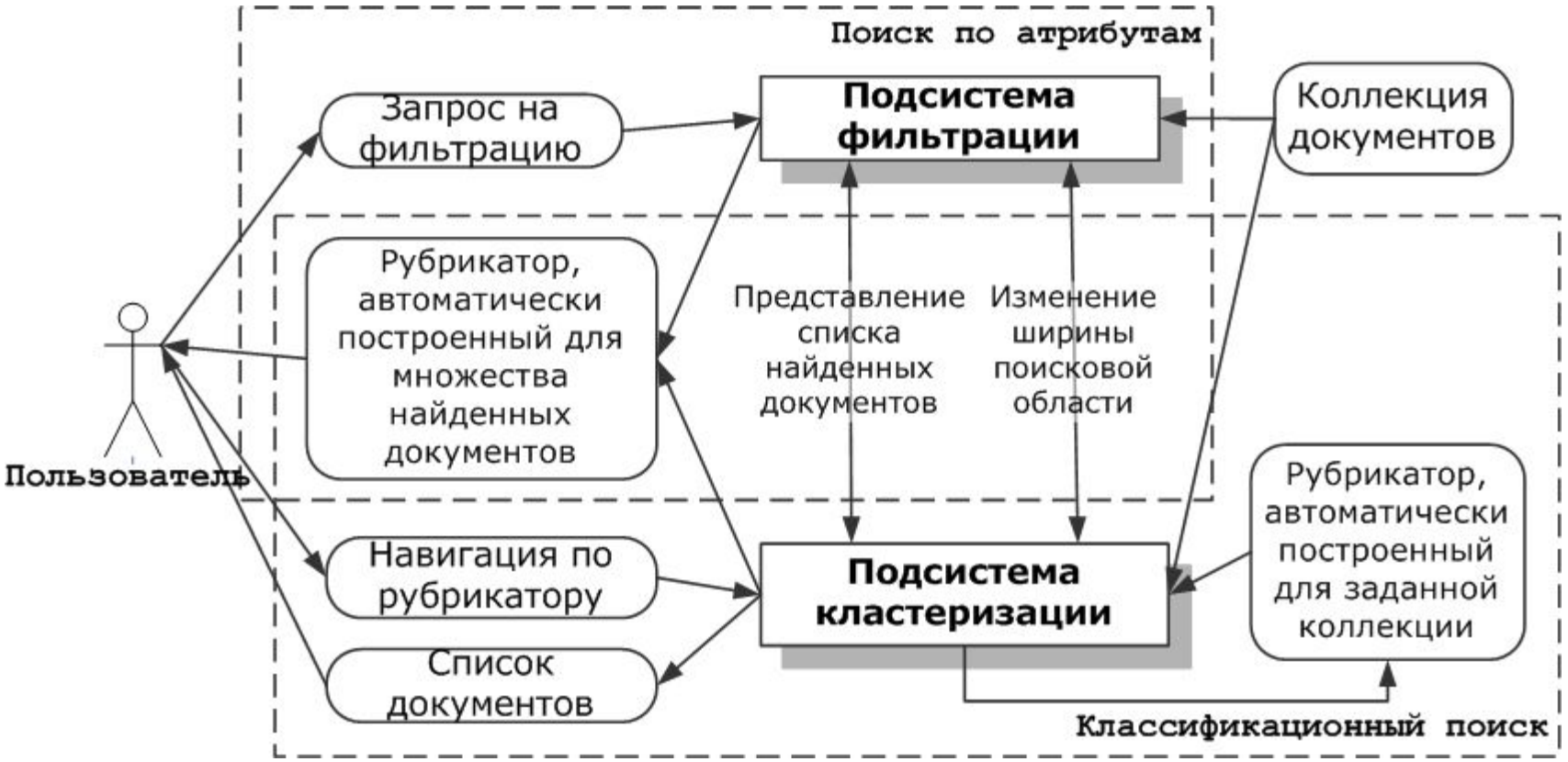
Московский государственный технический университет им. Н.Э.Баумана

2008

Постановка задачи

- Дано:
  - фонд полнотекстовых документов университетской библиотеки (учебные, обзорно-аналитические материалы различного объёма)
- Требуется:
  - создать средство тематической навигации по всему фонду или по его подмножествам, способное автоматически подстраиваться под тематику конкретного набора документов.

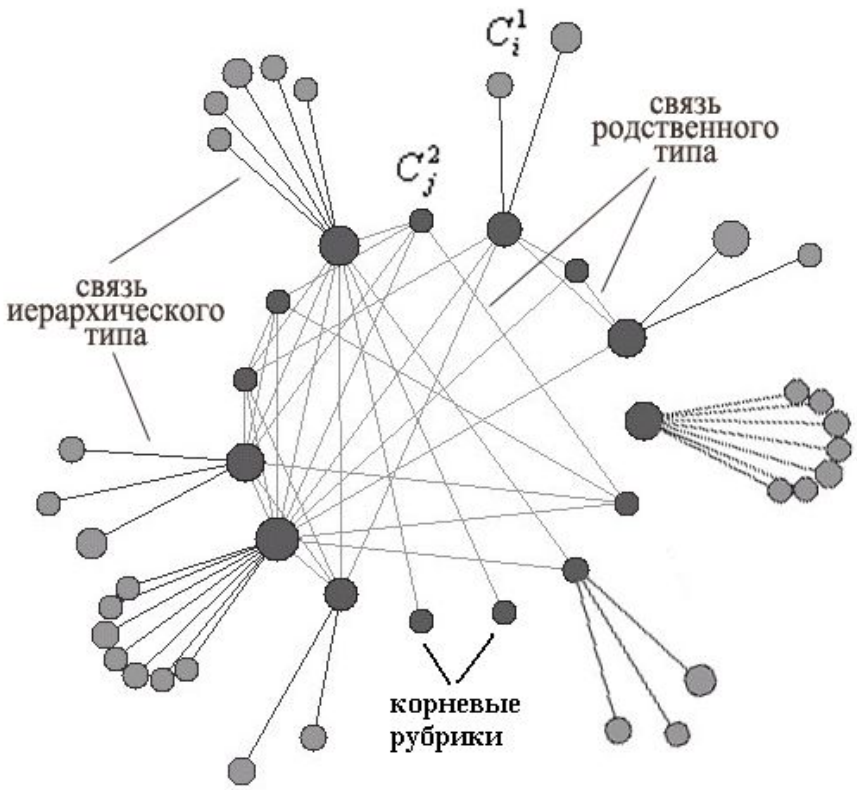
Механизм применения средства тематической навигации



Требования к виду искомой навигационной схемы

- Рубрикатор, унаследовавший основные характеристики от традиционного предметного рубрикатора библиотеки МГТУ им. Н. Э. Баумана:
  - иерархические связи между рубриками (не более 2-3 уровней);
  - родственные связи между рубриками (типа «см. также»);
  - краткое описание и список ключевых слов.

Способ представления рубрикатора



Рубрикатор в виде графа $G^* = (V^*, E^*)$, где

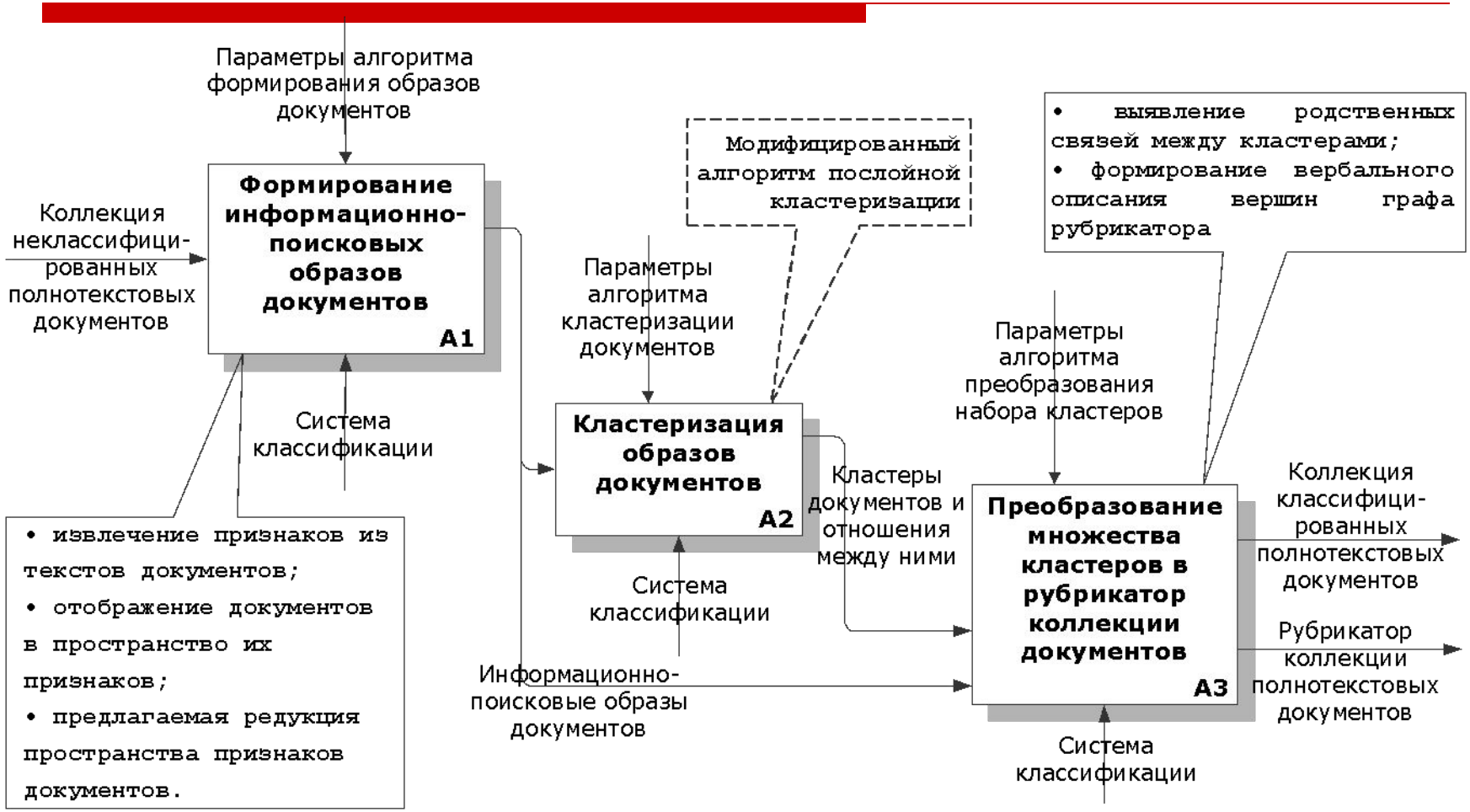
$V^* = \{v_1, ..., v_{N_C}\}$ – это множество вершин графа, отражающих кластеры документов, полученные при кластеризации коллекции полнотекстовых документов на заданном уровне иерархии;

$E^* = (e_{ij})$ – множество рёбер графа, отражающих как иерархические так и родственные связи.

Граф G^* является многоуровневым и содержит подграфы $G^0 \subseteq ... \subseteq G^t \subseteq ... \subseteq G^*$

Каждая выявленная группа документов должна иметь название и список ключевых слов.

Функциональная схема автоматического формирования рубрикатора

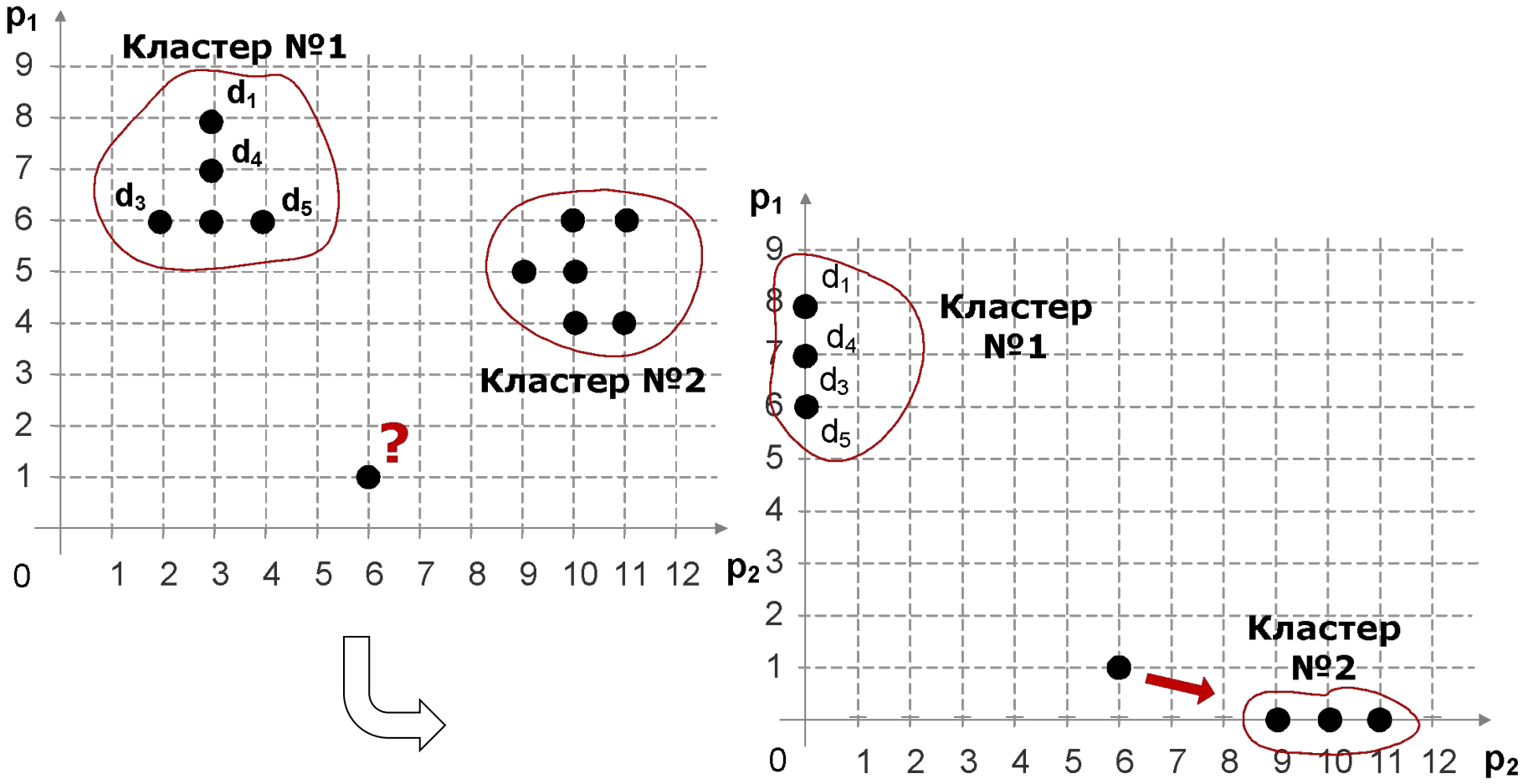


Выбор подхода к формированию образов документов

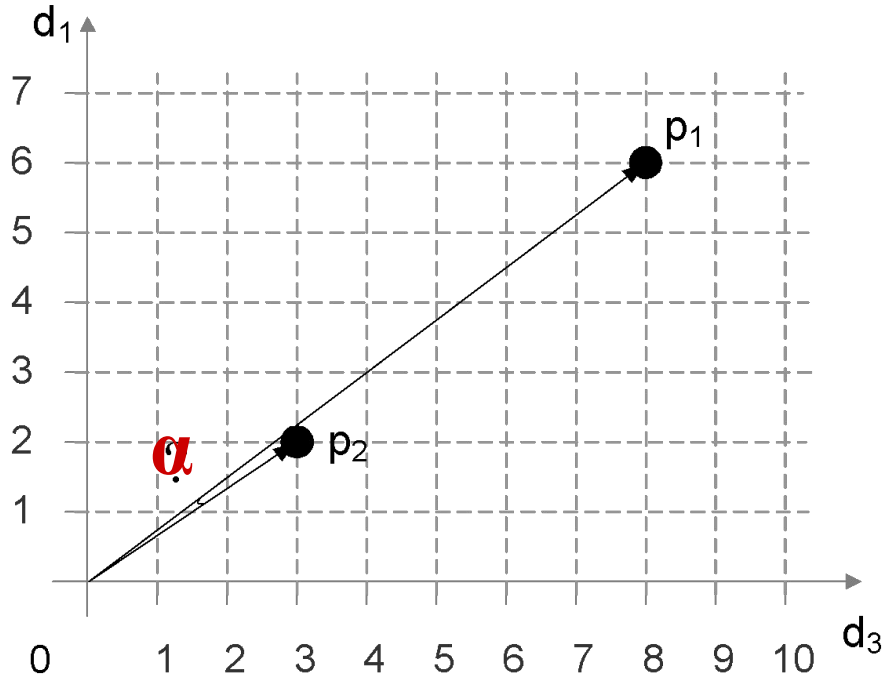
Предложенный алгоритм формирования образов документов

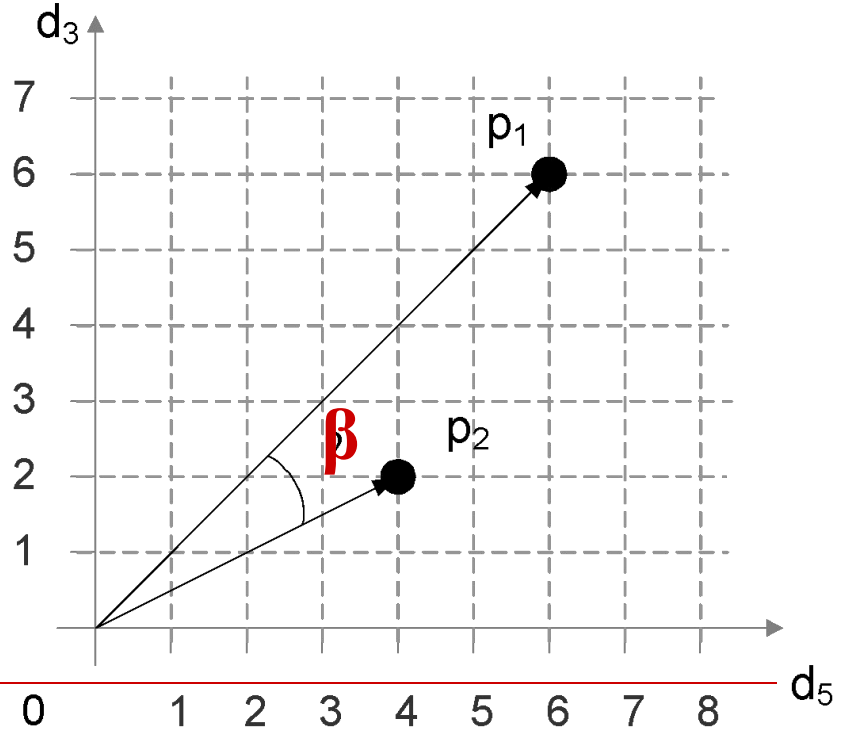
- Построение словаря признаков (одиночных слов) всех документов (морфологический анализ – стеммер М. Портера).
- Принудительная редукция признаков:
  - удаление стоп-слов;
  - Удаление слов по критерию документальной частоты с порогами t^{DF}_{min} и t^{DF}_{max}, где t^{DF}_{min}=<1 документ> и t^{DF}_{max}=<80% документов>.
- Взвешивание признаков документов по схеме TFIDF.
- Принудительная редукция признаков (продолжение):
  - для каждого документа в отдельности удаление некоторой доли t^{WP} самых маловесомых признаков, где t^{WP}=0.60.
- Избирательная редукция:
  - удаление из образов некоторых документов тех признаков, что обладают слабой различительной способностью для представления некоторого тематического класса.

Иллюстрации к предположению об избирательной редукции (1)



Иллюстрации к предположению об избирательной редукции (2)





Алгоритм избирательной редукции

- **выявить группы «очевидно родственных» документов $\{C_1^0,\ldots,C_{|c^0|}^0\}$ со значением меры близости документов не менее τ^{sim};**
- **для каждой группы C_i^0 удалить малоинформативные для её тематики признаки следующим образом:**

- перейти в подпространство документов данной группы $\prod^{C_i^0}(P)$:

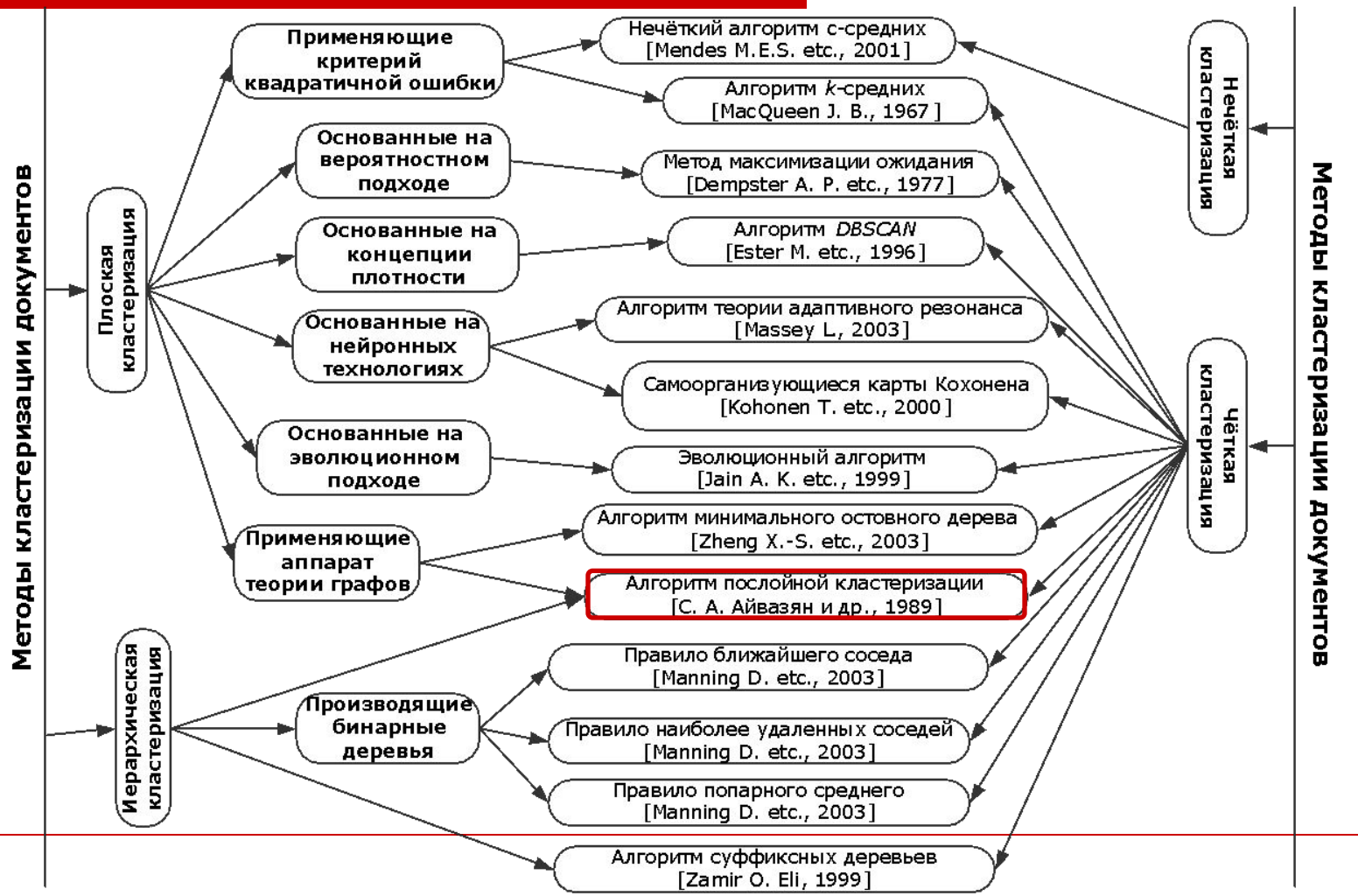
$$C_i^0 = (\vec{d}_1^{(i)},\ldots,\vec{d}_{N_i}^{(i)}) = \begin{pmatrix} w_{1(i)}^{(1)} & \cdots & w_{N_i(i)}^{(1)} \\ & \cdots & \\ w_{1(i)}^{(N_P)} & \cdots & w_{N_i(i)}^{(N_P)} \end{pmatrix} = \begin{pmatrix} \vec{p}_1 \\ \ldots \\ \vec{p}_{N_P} \end{pmatrix}$$

- в подпространстве документов $\prod^{C_i^0}(P)$ выполнить группировку признаков с максимальным значением меры близости внутри групп;

- удалить признаки, образовавшие группы в подпространстве документов, из образов $\vec{d}_j \in C_j^0$.

- удалить из множества *P* признаки, которые перестали принадлежать хотя бы одному $\vec{d}_j \in D$.

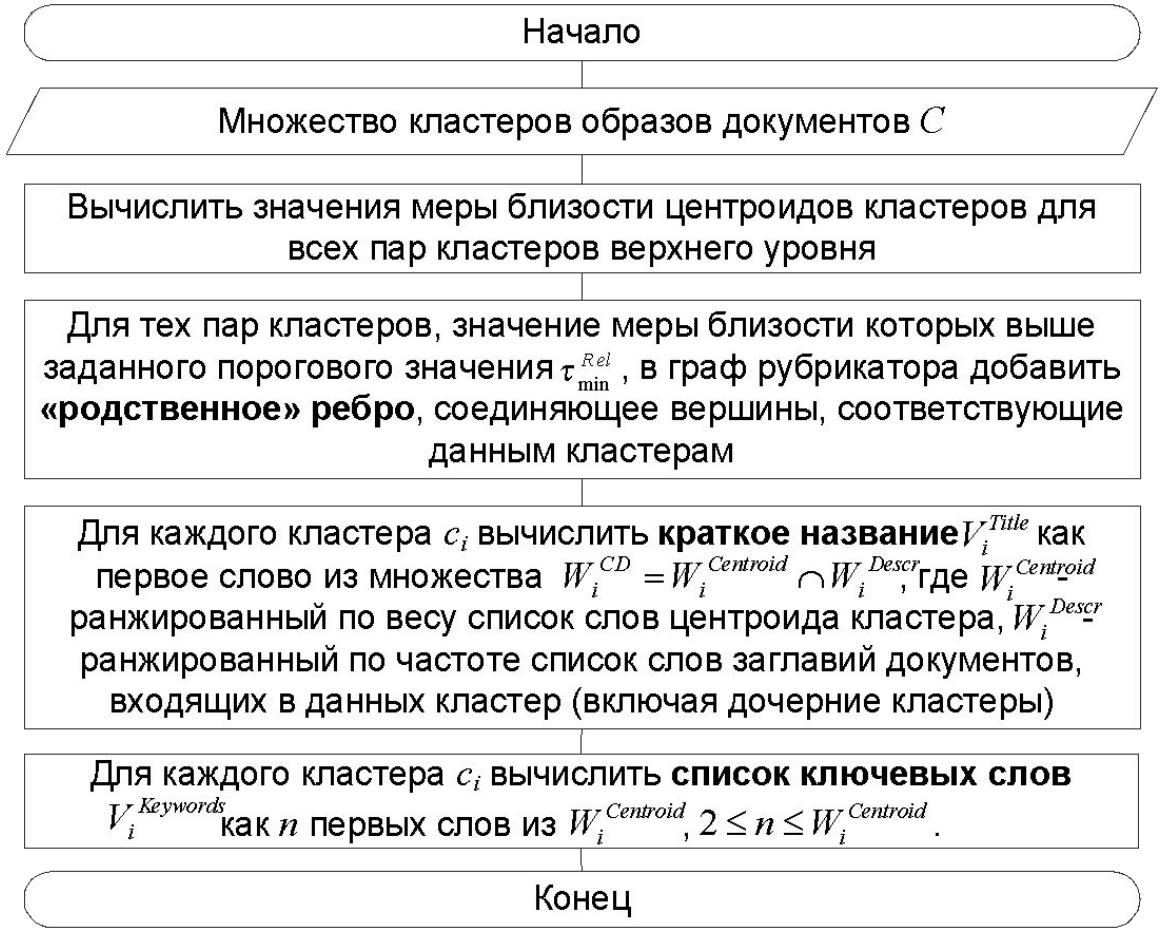
Выбор алгоритма кластеризации



Модифицированный алгоритм кластеризации документов

Результат: последовательность вложенных разбиений $C^1 \subset \ldots \subset C^m$, которые отражают иерархические связи между кластерами документов.

Дополнение кластерной структуры до искомого рубрикатора



Тестовые коллекции

1) **On-line библиотека CITFORUM (http://www.citforum.ru)**: наработка эмпирических сведений к методу формирования рубрикатора и оценка его эффективности **(CL1572)**.
2) **Ресурсы библиотеки МГТУ им. Н. Э. Баумана – коллекция авторефератов диссертаций** – апробация системы формирования рубрикатора **(TAL234)**.
3) **Коллекция нормативно-правовых документов** законодательства Российской Федерации, сформированная в 2004 году для выполнения заданий в рамках РОМИП **(Legal2004_5000)**. Отобраны те документы, для которых есть информация о их принадлежности рубрикам, - ~~25034~~ 5000 документов.

Меры качества кластеризации

- **Внешние меры**: автоматическое сравнение полученного разбиения документов с «эталонным» разбиение на кластеры (рубрики).
- **Внутренние меры**: автоматическая оценка свойств отделимости и компактности полученного разбиения документов.

Внешние меры качества кластеризации

- Полнота
- Точность
- F1-мера
- Ошибка
- Аккуратность
- и др.

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

Внутренние меры качества кластеризации

- Оценка иерархического разбиения:
  - Кофенетический коэффициент корреляции (CPCC)
- Оценка плоского разбиения:
  - Индекс Данна (Dunn, DI)
  - Индекс Девиса-Булдина (Davies-Bouldin, DB)
  - Индекс Калинского и Гарабача (Calinski и Harabasz, CH)
  - I-индекс (I-index)

Испытания алгоритма формирования образов (на CL1572)

	(1)	(2)	(3)
Внешние меры качества кластеризации			
MicroF$_1$-мера	0,190	0,466	**0,505**
Error	0,506	0,101	**0,084**
Внутренние меры качества кластеризации			
|CPCC|	0,188	0,508	**0,580**
DI	0,539	**0,598**	0,577
DB	0,196	0,258	**0,180**
CH	3,086	6,3687	**7,226**
I-Index	0,0014	0,0016756	**0,0018**
Скорость кластеризации			
Время (с)	1783	584	**85**

Оценка способа формирования образов.
(1) – без редукции,
(2) – с принудительной редукцией,
(3) – с принудительной и избирательной редукцией

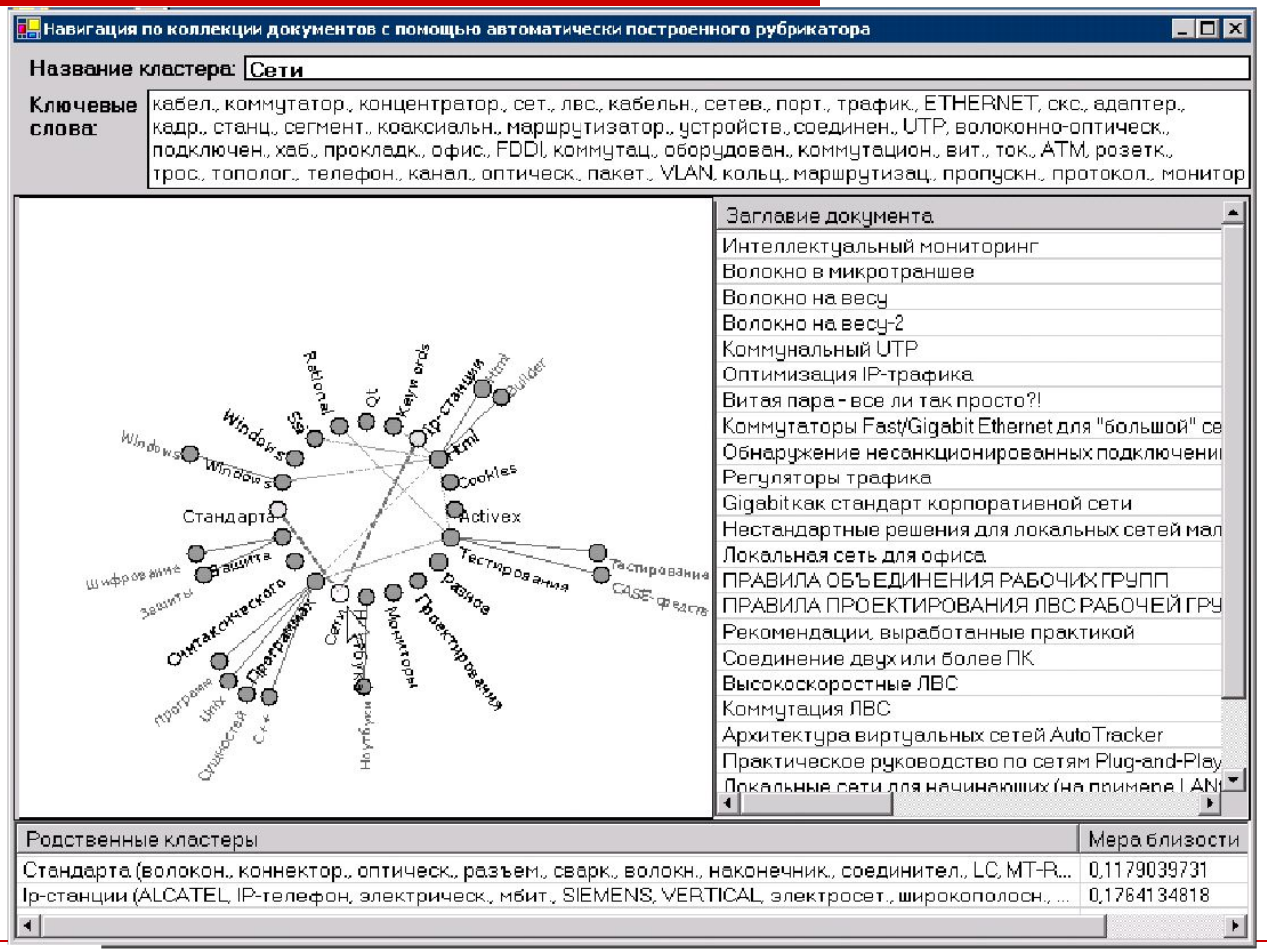
Испытание модифицированного алгоритма кластеризации (на CL1572)

	(1)	(2)	(3)
Внешние меры качества кластеризации			
MicroF$_1$-мера	0,21838	0,11321	**0,25823**
MacroF$_1$-мера	0,10998	0,06513	**0,14383**
Error	0,05741	0,30799	**0,03439**
Внутренние меры качества кластеризации			
CPCC	**-0,4642**	-0,1399	-0,3553
DI	**0,598**	0,378	0,500
DB	0,394	**0,053**	0,076
CH	**13,505**	2,746	6,278
I-Index	**0,000041**	0,000012	0,000022
Скорость кластеризации			
Время (с)	12503	1216	**624**

Оценка алгоритма кластеризации:

(1) – иерархический агломеративный алгоритм (усечение дерева при пороге меры близости – 0,20),
(2) – исходный алгоритм послойной кластеризации (два уровня при порогах меры близости {0,40; 0,20}),
(3) – модифицированный алгоритм послойной кластеризации (два уровня при порогах меры близости {0,40; 0,20}).

Пример интерфейса навигации по подмножеству CL1572



Пример интерфейса навигации по выборке из коллекции документов с помощью автоматически построенного рубрикатора (выбрана рубрика "Сети")

Испытание модифицированного алгоритма кластеризации (на TAL234)

- Ошибка автоматической классификации на TAL234:
  - 3,2% - в сравнении с классификацией авторефератов по УДК;
  - 13,6% - в сравнении с областью знания по номенклатуре ВАК , что объясняется тематическим перекрытием укрупнённых направлений, по которым осуществляется подготовка и защита диссертаций.

Испытания системы на Legal2004_5000 (1)

- Оценить качество кластеризации предложенным методом со значениями параметров, подобранными ранее на других коллекциях.
- Сравнить качество кластеризации при различных значениях параметров алгоритмов.
- Продолжить экспериментальное исследование алгоритма избирательной редукции.
- Оценить устойчивость метода (например, методом половинного деления).
- Оценить зависимость значений внешних и внутренних мер качества кластеризации.
- Усовершенствовать алгоритм формирования названий кластеров.

Испытания системы на Legal2004_5000 (2)

	(1)	(2)
Внешние меры качества кластеризации		
MicroF1-мера	**0,684**	0,683
MacroF1-мера	0,736	**0,739**
Error	0,390	**0,382**
Внутренние меры качества кластеризации		
|CPCC|	0,095	**0,102**
DI	**0,457**	0,456
DB	0,555	**0,426**
CH	2,881	**3,013**
I-Index	**0,0000134**	0,0000122
Характеристики пространства признаков		
Количество признаков	99 872 (37%)	97 869 (36%)
Количество связей типа «признак-документ»	2 418 482 (39%)	2 314 665 (37%)

Оценка кластеризации модифицированным алгоритмом (Legal2004_5000):
(1) – с принудительной редукцией,
(2) – с принудительной и избирательной редукцией (порог меры близости = 0,60).

Испытания системы на Legal2004_5000 (3)

Количественные характеристики пространства признаков

Пространство признаков	Число термов	Число связей типа «терм-документ»
Исходное	273 563	6 241 986
После принудительной редукции по критерию DF	107 122 (39%)	5 902 589 (95%)
После принудительной редукции по весу внутри отдельных документов (WP)	99 872 (37%)	2 418 482 (39%)
После избирательной редукции	97 869 (36%)	2 314 665 (37%)

Дальнейшие планы

- Закончить эксперименты на 5000 документов
- Провести исследования на 25034 документов
- Получить основания для выбора дальнейшего пути развития метода формирования рубрикатора

Вопросы

opeskova@mail.ru