



ТЕМАТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ НОВОСТНОГО КЛАСТЕРА КАК ОСНОВА ДЛЯ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ

Аспирант ВМК МГУ
Алексеев Алексей



ПЛАН ДОКЛАДА

2

- I. Задача автоматического аннотирования и словное представление текстов
- II. Тематический анализ новостных кластеров
- III. Интеграция тематического представления в методы аннотирования
- IV. Оценка автоматических аннотаций
- V. Заключение

АВТОМАТИЧЕСКОЕ АННОТИРОВАНИЕ И ПОСЛОВНОЕ ПРЕДСТАВЛЕНИЕ

3

- ❑ Новостные сервисы (30-40 тыс. документов в день)
- ❑ Кластеризация новостей на одну тему – новостной кластер
- ❑ Автоматическое аннотирование: отражение наиболее значимой информации текстовой коллекции в сжатом виде

Проблемы:

- ✓ Одна сущность названа посредством цепочки слов (многословным выражением)
- ✓ В кластере используется много разных наименований одной и той же сущности

Например: *Авиабаза США в Киргизии*

База Манас; авиабаза Манас; Манас; база США; американская авиабазабаза в международном аэропорту Манас; американская авиабаза; и т.д.

УСЛОЖНЕНИЕ ПОСЛОВНОГО ПРЕДСТАВЛЕНИЯ СУЩЕСТВУЮЩИЕ ПОДХОДЫ

4

- Использование информации, описанной в определенных ресурсах:

- ✓ *Английский язык: **Wordnet*** (Barzilay R., Elhadad M., 1999)
- ✓ *Русский язык: **PyTез*** (Лукашевич Н.В., Добров Б.В., 2009)

СПОРТИВНЫЙ ФИНАЛ	36
СПОРТИВНОЕ	54
СОРЕВНОВАНИЕ	
СПОРТ	8
СПОРТСМЕН	2
ФИНАЛИСТ	1

- **Latent Dirichlet Allocation (LDA)**

D. M. Blei, A. Y. Ng, M. I. Jordan, USA, 2003

- Вероятностная модель, разбивающая слова на тематические группы

- **Проблемы:**

- Вероятностный результат
- Сложность интерпретации результатов

“Arts”

NEW
FILM
SHOW
MUSIC
MOVIE
PLAY
MUSICAL
BEST
ACTOR
FIRST

“Children”

CHILDREN
WOMEN
PEOPLE
CHILD
YEARS
FAMILIES
WORK
PARENTS
SAYS
FAMILY

ТЕМАТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ

5

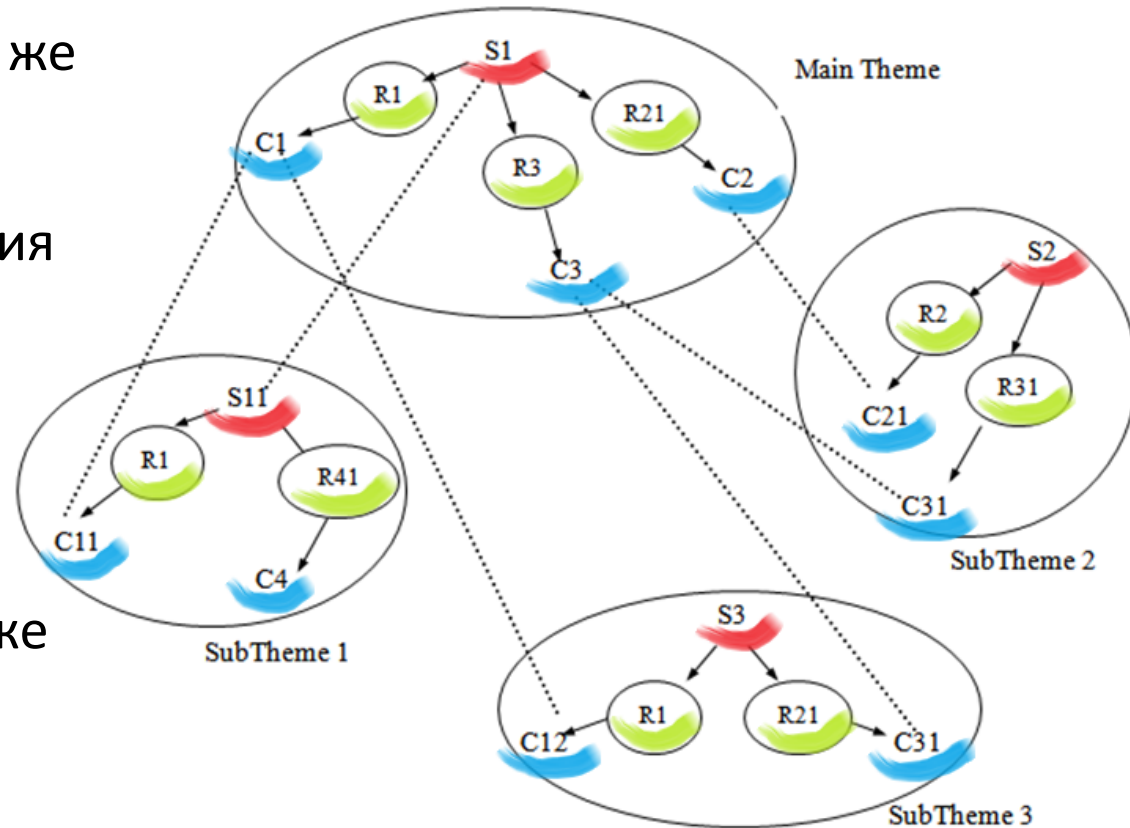
- ❑ Кластер документов должен соответствовать ситуации или ряду связанных ситуаций (*основная тема кластера*)
- ❑ В ситуации есть набор участников, которые в кластере:
 - Могут быть выражены словосочетаниями
 - Каждый участник может выражаться набором разных выражений
- ❑ **Тематическое представление:** совокупность основных участников ситуации (модель представления)
- ❑ Качество кластеров: ~70% (Добров Б.В, Павлов А.С. 2010) т.е. основные участники в среднем представлены правильно
- ❑ **Предположение:** более точное выявление основных участников ситуации поможет более правильно осуществлять разные операции с кластерами

ГЛАВНАЯ ТЕМА И ЕЕ УТОЧНЕНИЕ В КОНКРЕТНЫХ ПРЕДЛОЖЕНИЯХ

7

✓ Чем больше две сущности упоминаются в одних и тех же предложениях текста, тем более важно отношение между ними для содержания текста и **поэтому они относятся к разным участникам ситуации**

✓ Если сущности часто упоминаются, но мало встречаются в одних и тех же предложениях текста, то вероятно они имеют семантическую связь и **относятся к одним и тем же участникам ситуации**
(основано на Van Dijk, 1985)



ПОСТАНОВКА ЗАДАЧИ

8

❑ Входные данные:

Кластер тематически близких документов, собранных на основе пословной модели представления документов

❑ Задача:

Построить тематическое представление входного новостного кластера – совокупность основных участников ситуации и отношений между ними

❑ Применение:

- Добавление в пословное представление документа выделенных структур
- Использование усложненного представления документов кластера для улучшения качества операций с кластером (автоматического аннотирования)

ПРИНЦИПЫ ФОРМИРОВАНИЯ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

9

- ❑ Структуры, соответствующие участникам ситуации, собираются в форме **тематических узлов**:
 - центр и элементы ("похожие" на центр слова)
 - элементы могут принадлежать нескольким (двум) тематическим узлам
- ❑ Схожесть по нескольким характеристикам
- ❑ Если два выражения часто употребляются в одном предложении, они не могут оказаться в одном тематическом узле - ограничивающий фактор
- ❑ 4 типа контекстов: через глагол (**AV**); рядом (**Near**); не рядом (**NN**); в соседних предложениях (**NS**)
- ❑ Предварительная сборка многословных выражений

ИЗВЛЕЧЕНИЕ МНОГОСЛОВНЫХ ВЫРАЖЕНИЙ

10

- ❑ Рассматривается как необходимый этап для построения тематического представления
- ❑ Превышение встречаемости слов непосредственно рядом друг с другом по сравнению с отдельной встречаемостью во фрагментах предложений:

$$\text{Near} > 2 * (\text{Av} + \text{NotN})$$

- ❑ Дополнительные ограничения по частотности встречаемости слов рядом друг с другом
- ❑ **Примеры:** *Президент Алроса; президент компании; владелец компании; контрольный пакет акций; акции компании; пакет акций*

АЛГОРИТМ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

11

- Разные признаки сходства:
 - **Контекстно-независимые признаки** (формальное сходство, тезаурус)
 - **Контекстно-зависимые признаки** («строгие» контексты, статистика контекстов, вхождения в соседние предложения)
- **Итеративный алгоритм.** Одна итерация = объединение одной пары объектов
- **Управляющее правило:**
$$NS \geq 2 * (AV + Near + NN)$$

КОНТЕКСТНО-НЕЗАВИСИМЫЕ ПРИЗНАКИ

12

А. Формальное сходство объектов (BS)

- Простая метрика – одинаковые начала слов
- Пример: *Руководитель* – *Руководство*
- Вес признака: [0, 1]

$$BS = 1.0 - 0.1 * (\{ \text{Число_слов_с_отличными_началами} \})$$

В. Наличие связи в тезаурусе (TS)

- Рассмотрение цепочек отношений между объектами (использование вывода связей)
- Пример: *Руководитель* – *Глава*
- Вес признака: [0, 1]

$$TS = 1.0 - 0.2 * (\{ \text{Длина_пути_по_отношениям_тезауруса} \})$$

КОНТЕКСТНО-ЗАВИСИМЫЕ ПРИЗНАКИ

13

A. Частота появления в соседних предложениях (NSF)

$$NSF = \text{Min} \left[1, \frac{C}{\text{Avg}(C)} \right] \quad C = NS - 2 * (AcrossVerb + Near + NotNear)$$

B. Анализ «строгих» контекстов (SC)

- 4-граммы: по 2 слова влево и вправо
- Значение: $[0, 1]$ (относительно «лучшей» пары)

C. Скалярные произведения контекстов (SPS)

- Косинусная мера угла между векторами контекстов
- Сумма по характеристикам внутри предложения ($AV + Near + NotNear$)

ПРИМЕР РАНЖИРОВАНИЯ ПАР

14

Features Pairs	Context-independent		Context-dependent			SCORE
	BS	TS	NSF	SC	SPS	
Президент России – Президент РФ	0.90	1.00	0.00	0.50	0.68	2.74
Инвестгруппа– Инвестиционная группа	0.90	1.00	0.40	0.00	0.63	2.42
ГМК Норильский никель – Норильский никель	1.00	1.00	0.40	0.00	0.21	2.31
Российская Федерация – Россия	0.90	1.00	0.00	0.00	0.51	2.15
Отставка – Отставка с должности	0.90	1.00	0.40	0.00	0.00	2.10

ПРИМЕР СБОРКИ УЗЛА

15

- ❑ Итерация 7: (*Отставка*) $\leftarrow$ (*Отставка с должности*)
- ❑ Итерация 33: (*Отставка, Отставка с должности*) $\leftarrow$ (*Уход в отставку*)
- ❑ Итерация 44: (*Отставка, Отставка с должности, Уход в отставку*) $\leftarrow$ (*Отставка президента*)
- ❑ Итерация 61: (*Уход с поста*) $\leftarrow$ (*Уход в отставку*)
- ❑ Итерация 62: (*Отставка, Отставка с должности, Уход в отставку, Отставка президента*) $\leftarrow$ (*Уход с поста, Уход в отставку*)
- ❑ Итерация 102: (*Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста*) $\leftarrow$ (*Пост*)
- ❑ Итерация 103: (*Пост, Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста*) $\leftarrow$ (*Должность*)
- ❑ Итерация 104: (*Пост, Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста, Должность*) $\leftarrow$ (*Уход*)

РЕЗУЛЬТАТЫ РАБОТЫ АЛГОРИТМА НА КЛАСТЕРЕ ПРИМЕРЕ

16

- ❑ **Пост**: уход с поста; должность; уход; отставка; отставка с должности; уход в отставку; отставка президента
- ❑ **Алроса**: президент Алроса; АК Алроса
- ❑ **Компания**: акция компании; владелец компании; объединение компаний; акция; акционер компании; владелец; пакет акций; состав владельцев; контрольный пакет акций; контрольный пакет; владение
- ❑ **Ничипорук**: Александр Ничипорук
- ❑ **Якутия**: президент Якутии; якутский; якутский президент

ПРИМЕНЕНИЕ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

17

- Тематическое представление – дополнительная информация об устройстве новостного кластера

Слова → Объекты (слова + MB) → Тематические узлы

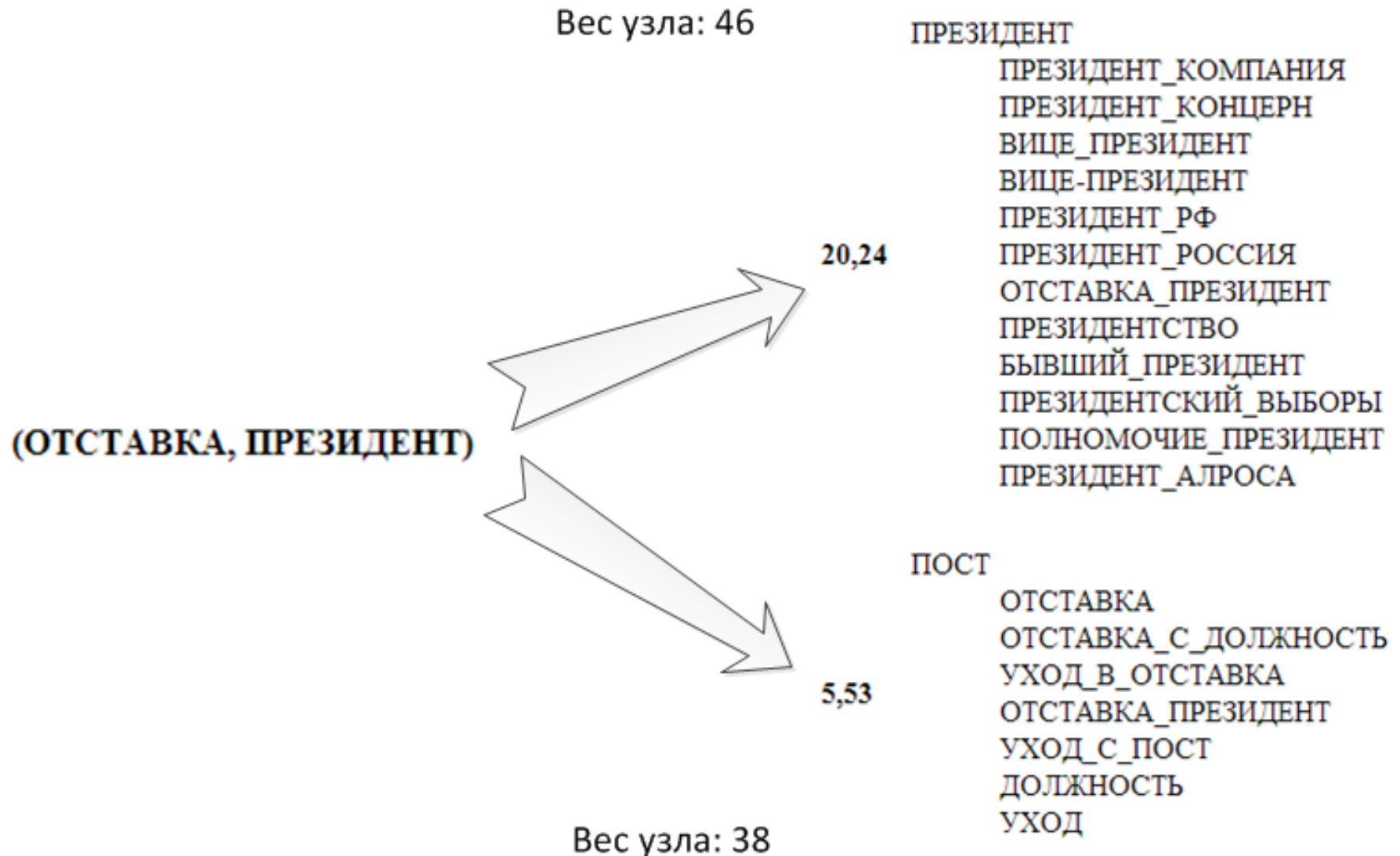
- Вес тематического узла:

$$weight(TU) = \sum_{ЭЛЕМЕНТ_ТУ_i \in TU} freq(ЭЛЕМЕНТ_ТУ_i)$$

- Вес объекта разбивается на веса в соответствующих тематических узлах (пропорционально силе связи с центром узла)

ПРИМЕНЕНИЕ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ - ПРИМЕР

18



ТЕМАТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ И АННОТИРОВАНИЕ

19

- Известные метода аннотирования:
 - Maximal Marginal Relevance (MMR)
 - SumBasic
 - Аннотирование на основе тем-го представления РуТез
- Специализированные методы аннотирования
 - Учет основных тематических узлов
 - Учет основных связей тематических узлов
- Учет **IDF**:

$$MWE_IDF = \ln \left(\frac{DOC_COUNT}{\left[\prod_{w_i \in MWE} Freq(w_i) \right] / [DOC_COUNT]^{N-1}} \right)$$

СОБСТВЕННЫЕ АЛГОРИТМЫ АННОТИРОВАНИЯ НА ОСНОВЕ ТЕМАТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ

20

Пусть: ТУ – тематический узел на основе предложенного алгоритма
ТУ_REL – пара ТУ, * NEW – новый ТУ или ТУ_REL

Предлагается два алгоритма аннотирования:

I. Отбор по тематическим узлам (*OurSummary_Nodes*)

$$\text{➤ } s_i \Rightarrow \max \left(\begin{array}{c} \text{desc weight}(TY_NEW_j) \\ \sum_{TY_NEW_j \in s_i, i=1..3} \text{weight}(TY_NEW_j) \end{array} \right)$$

II. Отбор по связям ТУ (*OurSummary_Relations*)

➤ $\text{weight}(TY_REL)$ – число вхождений в одни и те же предложения

$$\text{➤ } s_i \supset \max_{TY_REL_NEW_j \in Cluster} \left(\text{weight}(TY_REL_NEW_j) \right)$$

$$\text{➤ } s_i \Rightarrow \max \left(\sum_{TY_REL_NEW_j \in s_i} \text{weight}(TY_REL_NEW_j) \right)$$

ПРИМЕР АННОТАЦИИ

Президент "АЛРОСА" Александр Ничипорук уходит со своего поста.

1. Отношения между нынешним и бывшим президентами АЛРОСА (господин Штыров возглавлял компанию до того как стал президентом Якутии) не сложились с момента прихода господина Ничипорука два года назад.

2. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

ПРИМЕР АННОТАЦИИ

22

Президент "АЛРОСА" Александр Ничипорук уходит со своего поста.

1. Отношения между нынешним и бывшим президентами АЛРОСА (господин Штыров возглавлял компанию до того как стал президентом Якутии) не сложились с момента прихода господина Ничипорука два года назад.

2. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

ПРИМЕР АННОТАЦИИ

23

Президент "АЛРОСА" Александр Ничипорук уходит со своего поста.

1. Отношения между нынешним и бывшим президентами АЛРОСА (господин Штыров возглавлял компанию до того как стал президентом Якутии) не сложились с момента прихода господина Ничипорука два года назад.

2. Александр Ничипорук покидает пост президента компании АЛРОСА, на которую приходится четверть мировой добычи алмазов.

3. Акционерами "АЛРОСА" являются Росимущество - 37 % акций, Министерство по управлению госимуществом Якутии - 32 %, физические и юридические лица - 23 %.

АЛГОРИТМ ТЕСТИРОВАНИЯ

24

- ❑ Оценка результата нетривиальна: высокая степень субъективности и низкая согласованность экспертов
- ❑ 11 новостных кластеров, 2-4 ручные аннотации к каждому
- ❑ Оценка содержания аннотаций методом ROUGE (Chin-Yew Lin, 2004)
- ❑ Дополнительная оценка лучших алгоритмов методом «Пирамид» (Harnly A., Nenkova A., Passonneau R., 2005)
- ❑ Различные алгоритмы аннотирования в различных модификациях:
 - Maximal Marginal Relevance (MMR, 4 модификации)
 - SumBasic (2 модификации)
 - Собственный алгоритм аннотирования (4 модификации)
 - Аннотирование на основе тем-го представления PyTез

РЕЗУЛЬТАТЫ (ROUGE)

25

Метод	1	2	L	S	SU	Avg
MMR + Groups	0,62499 (1)	0,41633 (1)	0,6021 (1)	0,35529 (1)	0,36649 (1)	1,0
OurSummary_Nodes	0,58652 (2)	0,36154 (3)	0,5645 (2)	0,32113 (2)	0,33203 (2)	2,2
OurSummary_Nodes_WithIDF	0,58497 (3)	0,33918 (5)	0,55745 (3)	0,30124 (3)	0,31283 (3)	3,4
MMR_WithIDF	0,57623 (4)	0,38116 (2)	0,55503 (4)	0,29792 (4)	0,30971 (4)	3,6
MMR_WithoutIDF	0,56784 (5)	0,34595 (4)	0,55124 (5)	0,26092 (6)	0,27349 (6)	5,2
ThematicLines	0,53416 (6)	0,33364 (6)	0,51238 (6)	0,2713 (5)	0,28243 (5)	5,6
OurSummary_Relations	0,53141 (7)	0,2892 (7)	0,50422 (7)	0,25382 (7)	0,26509 (7)	7,0
SumBasic + Groups	0,52255 (8)	0,22881 (10)	0,493 (9)	0,24356 (8)	0,25525 (8)	8,6
SumBasic	0,51847 (9)	0,24735 (9)	0,49786 (8)	0,23064 (9)	0,24257 (9)	8,8
OurSummary_Relations_WithIDF	0,45494 (10)	0,24856 (8)	0,43768 (10)	0,19419 (11)	0,20492 (11)	10,0
MMR_WithIDF + Groups	0,44475 (11)	0,22238 (11)	0,42318 (11)	0,20627 (10)	0,21648 (10)	10,6

РЕЗУЛЬТАТЫ (ROUGE)

26

Метод	1	2	L	S	SU	Avg
MMR + Groups	0,62499 (1)	0,41633 (1)	0,6021 (1)	0,35529 (1)	0,36649 (1)	1,0
OurSummary_Nodes	0,58652 (2)	0,36154 (3)	0,5645 (2)	0,32113 (2)	0,33203 (2)	2,2
OurSummary_Nodes_WithIDF	0,58497 (3)	0,33918 (5)	0,55745 (3)	0,30124 (3)	0,31283 (3)	3,4
MMR_WithIDF	0,57623 (4)	0,38116 (2)	0,55503 (4)	0,29792 (4)	0,30971 (4)	3,6
MMR_WithoutIDF	0,56784 (5)	0,34595 (4)	0,55124 (5)	0,26092 (6)	0,27349 (6)	5,2
ThematicLines	0,53416 (6)	0,33364 (6)	0,51238 (6)	0,2713 (5)	0,28243 (5)	5,6
OurSummary_Relations	0,53141 (7)	0,2892 (7)	0,50422 (7)	0,25382 (7)	0,26509 (7)	7,0
SumBasic + Groups	0,52255 (8)	0,22881 (10)	0,493 (9)	0,24356 (8)	0,25525 (8)	8,6
SumBasic	0,51847 (9)	0,24735 (9)	0,49786 (8)	0,23064 (9)	0,24257 (9)	8,8
OurSummary_Relations_WithIDF	0,45494 (10)	0,24856 (8)	0,43768 (10)	0,19419 (11)	0,20492 (11)	10,0
MMR_WithIDF + Groups	0,44475 (11)	0,22238 (11)	0,42318 (11)	0,20627 (10)	0,21648 (10)	10,6

РЕЗУЛЬТАТЫ (ROUGE)

27

Метод	1	2	L	S	SU	Avg
MMR + Groups	0,62499 (1)	0,41633 (1)	0,6021 (1)	0,35529 (1)	0,36649 (1)	1,0
OurSummary_Nodes	0,58652 (2)	0,36154 (3)	0,5645 (2)	0,32113 (2)	0,33203 (2)	2,2
OurSummary_Nodes_ithIDF	0,58497 (3)	0,33918 (5)	0,55745 (3)	0,30124 (3)	0,31283 (3)	3,4
MMR_WithIDF	0,57623 (4)	0,38116 (2)	0,55503 (4)	0,29792 (4)	0,30971 (4)	3,6
MMR_WithoutIDF	0,56784 (5)	0,34595 (4)	0,55124 (5)	0,26092 (6)	0,27349 (6)	5,2
ThematicLines	0,53416 (6)	0,33364 (6)	0,51238 (6)	0,2713 (5)	0,28243 (5)	5,6
OurSummary_Relations	0,53141 (7)	0,2892 (7)	0,50422 (7)	0,25382 (7)	0,26509 (7)	7,0
SumBasic + Groups	0,52255 (8)	0,22881 (10)	0,493 (9)	0,24356 (8)	0,25525 (8)	8,6
SumBasic	0,51847 (9)	0,24735 (9)	0,49786 (8)	0,23064 (9)	0,24257 (9)	8,8
OurSummary_Relations_WithIDF	0,45494 (10)	0,24856 (8)	0,43768 (10)	0,19419 (11)	0,20492 (11)	10,0
MMR_WithIDF + Groups	0,44475 (11)	0,22238 (11)	0,42318 (11)	0,20627 (10)	0,21648 (10)	10,6

РЕЗУЛЬТАТЫ (ПИРАМИДЫ)

28

- Дополнительная оценка методом «Пирамиды» для алгоритмов аннотирования, показавших улучшение при оценке методом ROUGE

Метод	Score
MMR + Groups	0,645 (1)
MMR_WithIDF	0,617 (2)
OurSummary_Nodes	0,602 (3)
SumBasic + Groups	0,575 (4)
SumBasic	0,567 (5)

ЗАКЛЮЧЕНИЕ

29

- ❑ Предложена и реализована модель автоматического построения тематического представления новостного кластера, основанная на совокупности нескольких факторов
- ❑ Предложен метод автоматического аннотирования новостных кластеров на основе тематического представления
- ❑ Предложены методы интеграции тематического представления в алгоритмы автоматического аннотирования новостного кластера
- ❑ Показано улучшение качества работы алгоритмов аннотирования на основе тематического представления

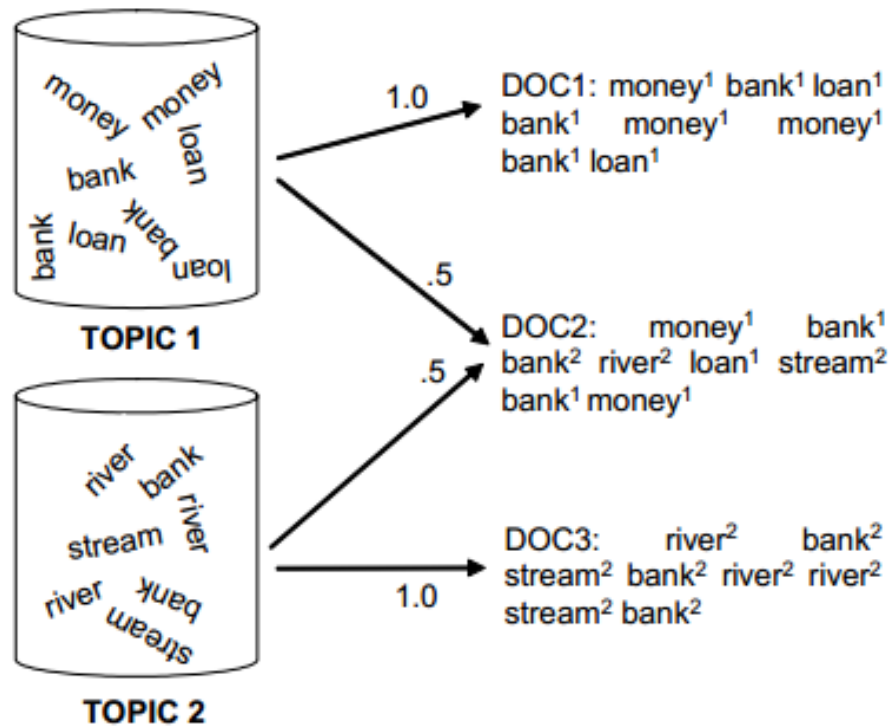
ДОПОЛНИТЕЛЬНЫЕ СЛАЙДЫ

Latent Dirichlet Allocation (LDA)

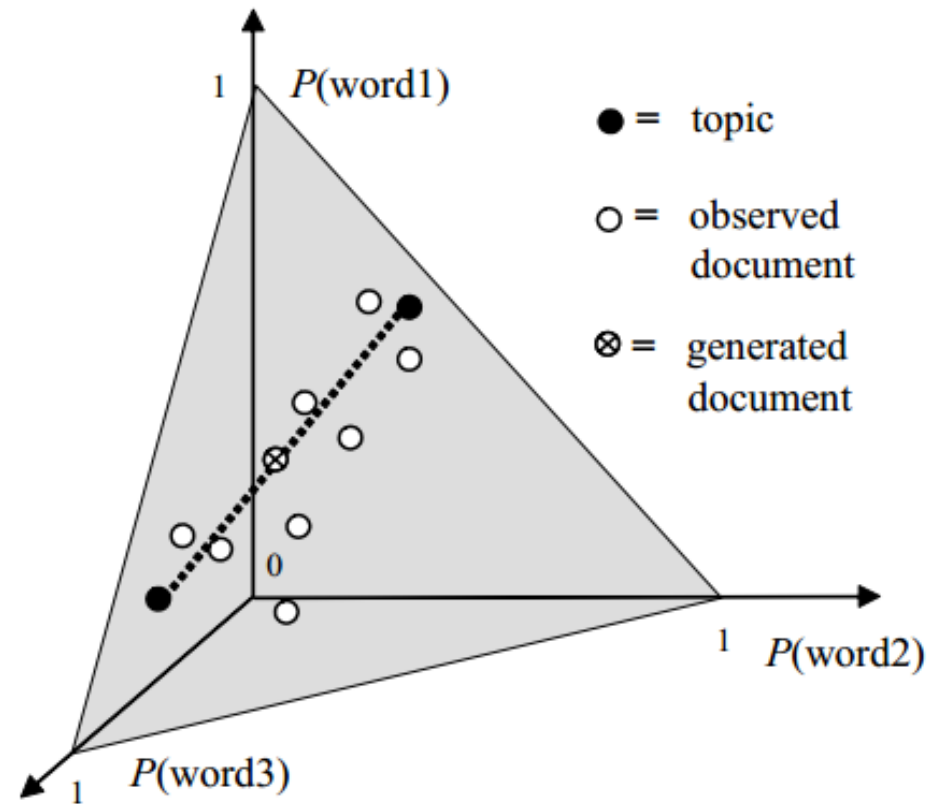
University of California, USA, 2003

David M. Blei, Andrew Y. Ng, Michael I. Jordan

31



Пример автоматического семплинга



Геометрическая интерпретация LDA

РАЗЛИЧИЯ С МОДЕЛЬЮ LDA

32

LDA	НАША МОДЕЛЬ
1. Совместная встречаемость => описание ситуации, сферы деятельности	1. Выделение различного именования основных участников
2. Результат разбиения – непрерывный (ненулевые вероятности)	2. Дискретное (строгое) разбиение
3. Не использует никаких лингвистических знаний	3. Ориентирована на аккумуляцию всех типов информации
4. Нет поддержки многословных выражений	4. Изначальное построение модели с учетом многословных выражений

МЕТОД ROUGE

33

- ❑ ROUGE - Recall-Oriented Understudy for Gisting Evaluation ()
- ❑ **Основная идея:** сопоставление порожденных аннотаций с ручными (составленными экспертами)
- ❑ Различные метрики сравнения: N-граммы, Longest Common Substring и их модификации

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

МЕТОД «ПИРАМИДЫ»

(Harnly A., Nenkova A., Passonneau R., 2005)

34

- Эксперты выделяют из «эталонных» аннотаций «информационные единицы» - Summary Content Units (SCUs).
- Каждый SCU получает вес, равный количеству «эталонных» аннотаций, где она встречалась.
- Оценка – суммарный вес входящих SCU.
- Неоднократное вхождение SCU в автоматическую аннотацию не поощряется.
- **Итоговый результат:**

$$\frac{[\text{Суммарный _ вес _ найденных _ SCU}]}{[\text{Суммарный _ вес _ всех _ SCU _ для _ данного _ топика}]}$$