Аннотированные суффиксные деревья как средство интерпретации текстов

Екатерина Черняк

Модели представления текста

- Векторная модель
- Языковая модель
- Модели скрытых тем
- Модель суффиксных деревьев

Векторная модель (Salton & Buckley, 1998)

- Текст – это множество термов $T=(t_1,\ldots, t_n)$
- Каждому терму соответствует своя координата в векторном пространстве
- Представление текста – вектор в пространстве термов, компоненты которого – частоты или какие-нибудь веса
- $d = (f_1,\ldots, f_n)$ или $d = (w_1,\ldots, w_n)$

Веса термов в векторной модели

- В (Salton & Buckley, 1998) предложена общая схема веса терма: $w_{ij} = l_{ij} * g_i$, где l – локальный вес, g – глобальный

- Некоторые варианты локальных весов (Berry & Browne, 2005):

  - Бинарный вес: $l_{ij} = 1$, если терм встречается в тексте, 0 в обратном случае

  - Частота: $l_{ij} = tf_{ij}$

  - Логарифмический вес: $l_{ij} = log(tf_{ij}+1)$

- Некоторые варианты глобальных весов
  - Бинарный вес: $g_j=1$
  - idf (inverse document frequency):$g_j = log\ (N/1+df_j)$
  - gf.idf (general frequence / idf) $g_j = gf_j/df_j$

- Самая популярная схема взвешивания – *tf-idf*: $w_{ij}=tf_{ij}* log\ (N/1+df_i)$

Достоинства векторной модели

- Простота построения по заданному корпусу текстов
- Использования линейно-алгебраических операций для определения сходства между текстами и поиска по запросу

Недостатки векторной модели

- Гипотеза о независимости термов
- Не учитываются кореферентные, анафорические, синонимические и другие связи между словами

Приложения векторной модели

- Категоризация текстов (Sebastiani, 2002)
- Классификация текстов, в т.ч. по тональности (Turney P. D., 2002) **и** (Pang, Lee, & Vaithyanathan)
- Кластеризация текстов (Andrews & Fox, 2007)

Развитие векторной модели

- Обобщенная векторная модель (generalized vector space model, GVSM) (Wong, Ziarko, & Wong, 1985)

- Векторные модели семантики (Pantel & Lin, 2002)

- Вероятностная модель релевантности (Robertson & Zaragoza, 2009)

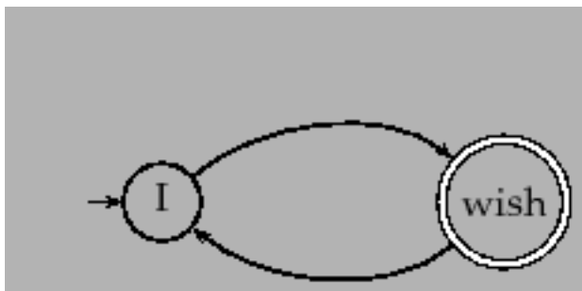
- Latent semantic indexing / analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

Языковая модель (Ponte & Croft, 1998)

- Оценивается вероятность того, что одно слово после предыдущего (модель униграм), после двух предыдущих слов (модель биграм), *n* слов (модель *n*-грам).

- Модель униграмм: $P(t_{1,n}) = P(t_1, t_2, \ldots, t_n) = \prod_i P(t_i)$

- Модель биграмм $P(t_i|t_{i-1}, t_{i-2}, \ldots, t_1) = \prod_i P(t_i|t_{i-1})$



I wish
I wish I wish
I wish I wish I wish
I wish I wish I wish I wish I wish I wish
…

CANNOT GENERATE: wish I wish

► **Figure 12.1** A simple finite automaton and some of the strings in the language it generates. → shows the start state of the automaton and a double circle indicates a (possible) finishing state.

(Manning, Raghavan, & Schütze, An Introduction to Information Retrieval, 2009)

Приложения языковой модели

- Поскольку языковая модель является генеративной, используется в задачах
  - машинного перевода (Koehn, Och, & Marcu, 2003)
  - распознавания речи (Katz, 1987)
  - исправления опечаток (Байтин, 2008)
- Может быть использования для оценивания вероятности запроса в тексте (Ponte & Croft, 1998)

Модели скрытых тем

- LSI/LSA, pLSI, LDA, LLDA, PAM и т.д.7
- Модели представления коллекций текстов
- Текст – это набор скрытых тем
- Тема состоит из слов

Latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

- Сингулярное разложение матрицы терм–текст
- Исходная матрица $X_{t\times d} = U_{t\times n}\Sigma_{n\times n}(V_{n\times d})^T$
- Снижение размерности $\widehat{X_{t\times d}} = U_{t\times k}\Sigma_{k\times k}(V_{k\times d})^T$

Latent semantic analysis

- Недостатки:
  - вычислительная сложность
  - непонятна природа тем
- Приложения:
  - Поиск по запросу (Wei & Croft, 2006)
  - Классификация текстов (Hyunsoo, Howland, & Park, 2005)
  - Фильтрация спама (Gee, 2004)
  - Суммаризация текста (Gong & Liu, 2001)
- Развитие модели: pLSI

Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003)

- Генеративная модель порождения коллекции текстов

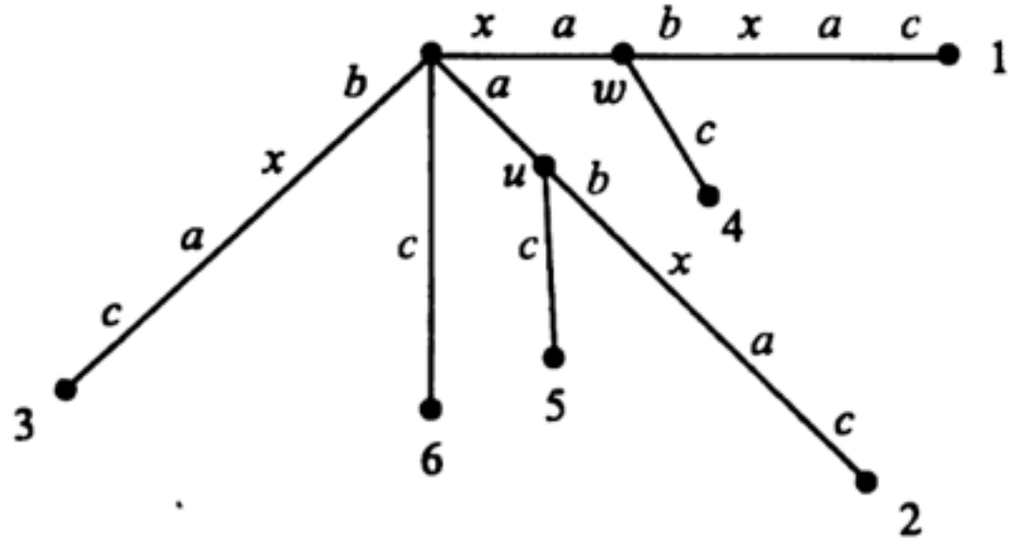
1. Пусть распределение тем в тексте i : $\theta_i \sim Dir(\alpha), 1 \leq i \leq M, Dir(\alpha)$ – распределение Дирихле с параметром α.
2. Пусть распределение слов в теме k : $\varphi_k \sim Dir(\beta), 1 \leq k \leq K$–число тем.
3. Для каждого позиции слова $i, j,\ 1 \leq i \leq M, 1 \leq j \leq M$:
    a. Выбрать тему $z_{ij} \sim Multinomial(\theta_i)$
    b. Выбрать слово $w_{ij} \sim Multinomial(\varphi_{z_{ij}})$.

Latent Dirichlet Allocation

- Недостатки: сложность восприятия
- Применение: почти везде
- Развитие: LLDA (Ramage, Hall, Nallapati, & Manning, 2009), PAM (McCallum, 2006)

Модель суффиксных деревьев (Weiner, 1973)

- Средство хранение строковых данных и поиска подстрок (но занимает много памяти)
- Ребра помечены суффиксами строки
- У каждого узла как минимум два потомка
- Листья пронумерованы



Кластеризация на основе суффиксных деревьев STC (Zamir, Etzioni, 1997)

- СД постреено по словам для коллекции текстов
- Каждый лист – это базовый кластер из номеров текстов
- Кластер – компонента связности в графе близости базовых кластеров

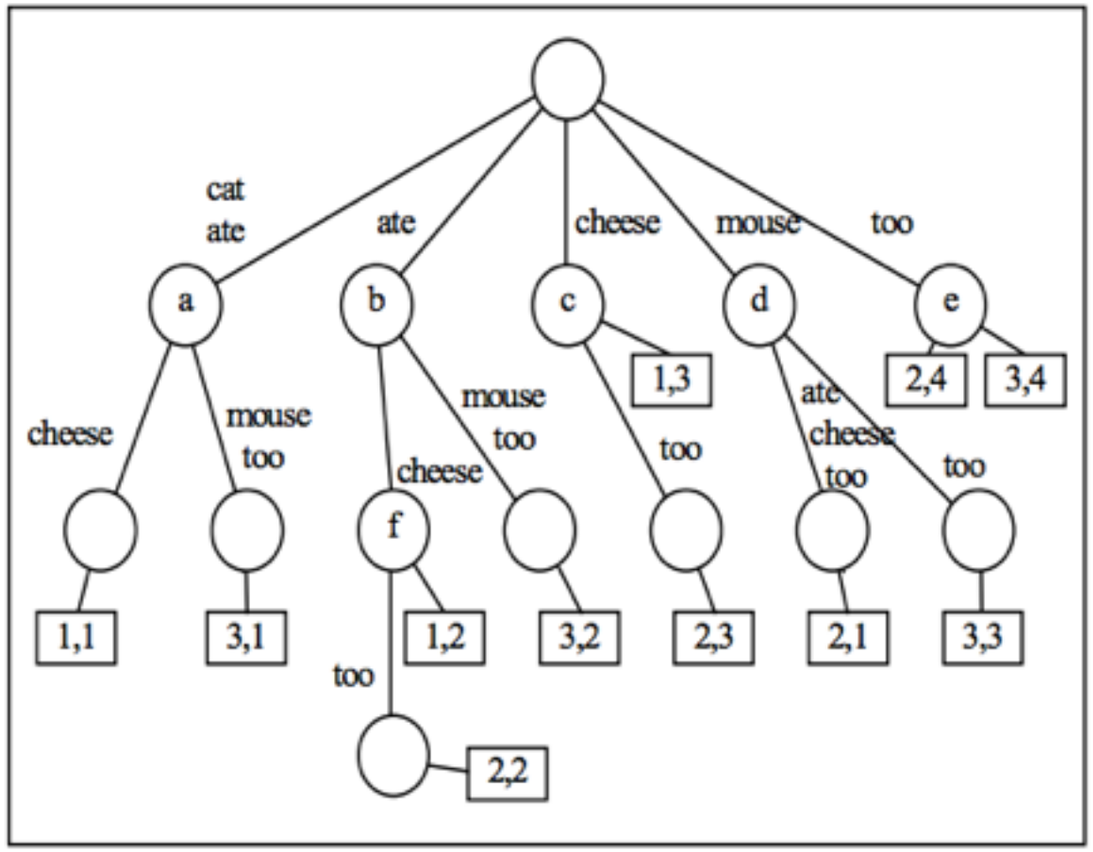
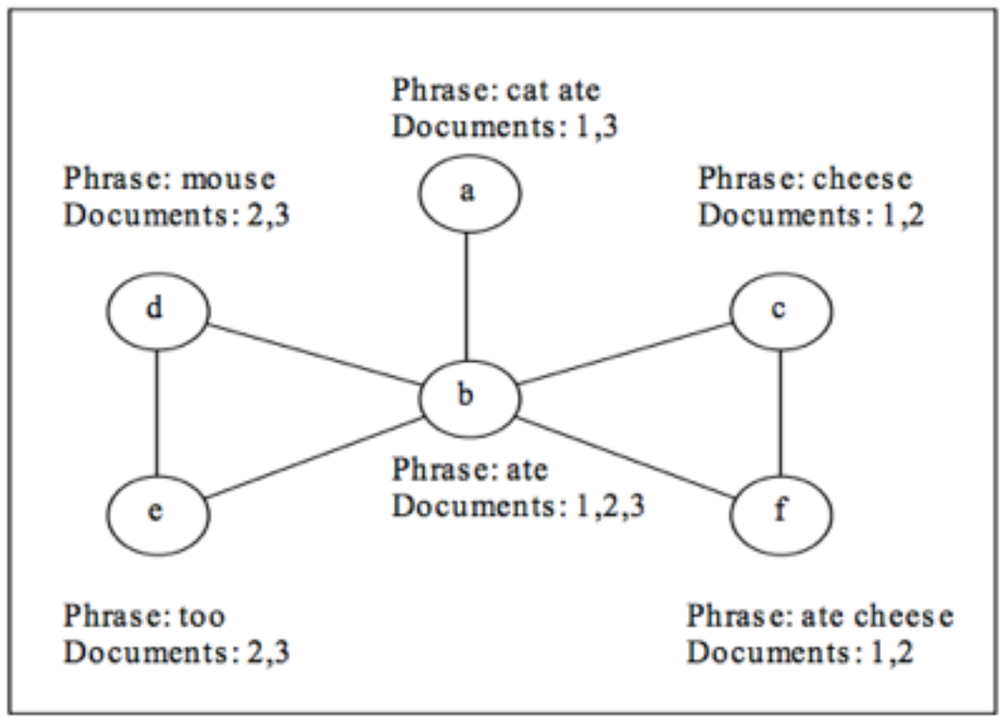


Figure 2: The suffix tree of the strings "cat ate cheese", "mouse ate cheese too" and "cat ate mouse too".



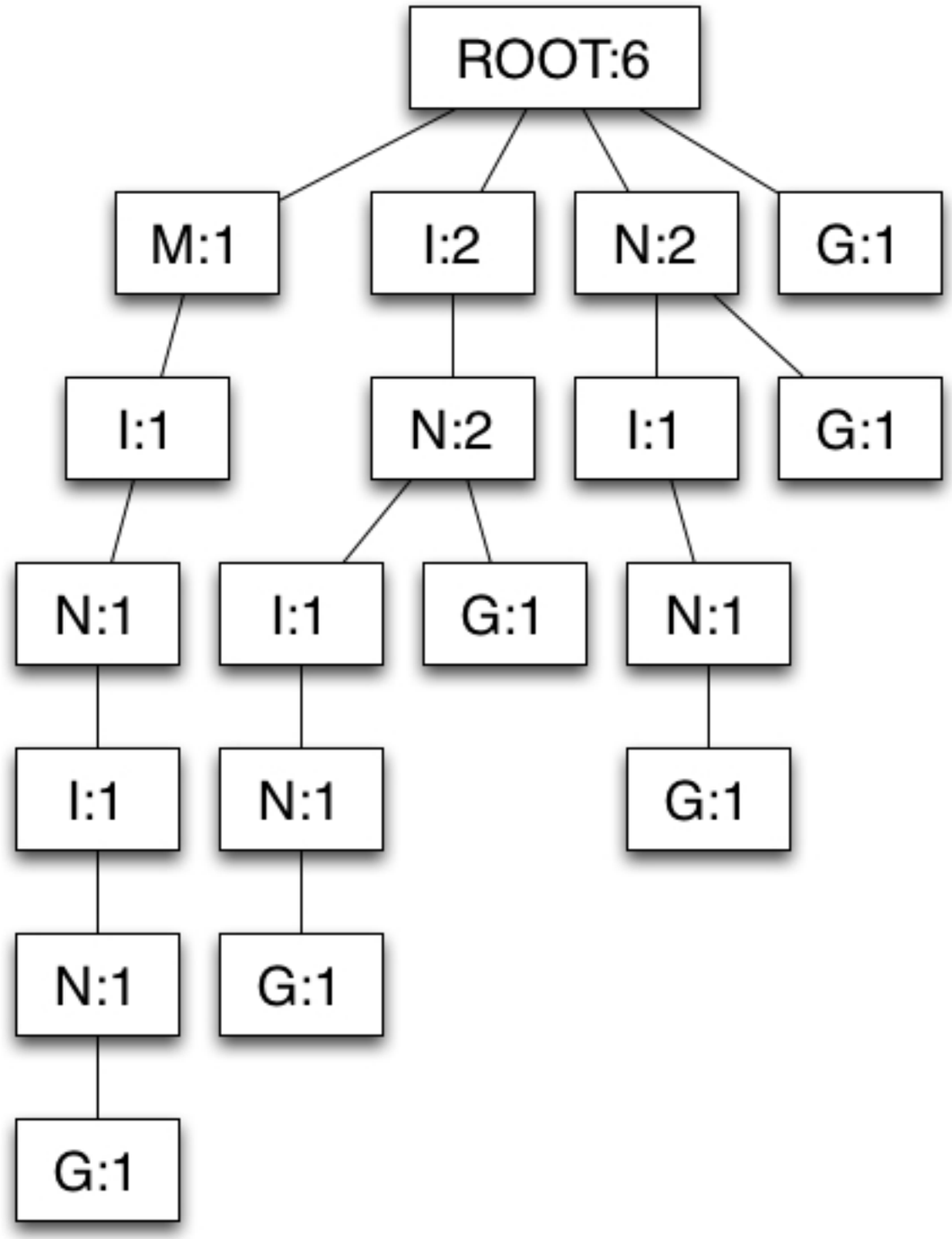
Другие приложения

- Выделение ключевых словосочетаний произвольной длины в потоках текстовых данных (Snowsill, Nicart, Stefani, Bie & Cristianini, 2010)

- Выделение именованных сущностей и жаргонизмов (Hu, Zhang, & Zhou, 2007)

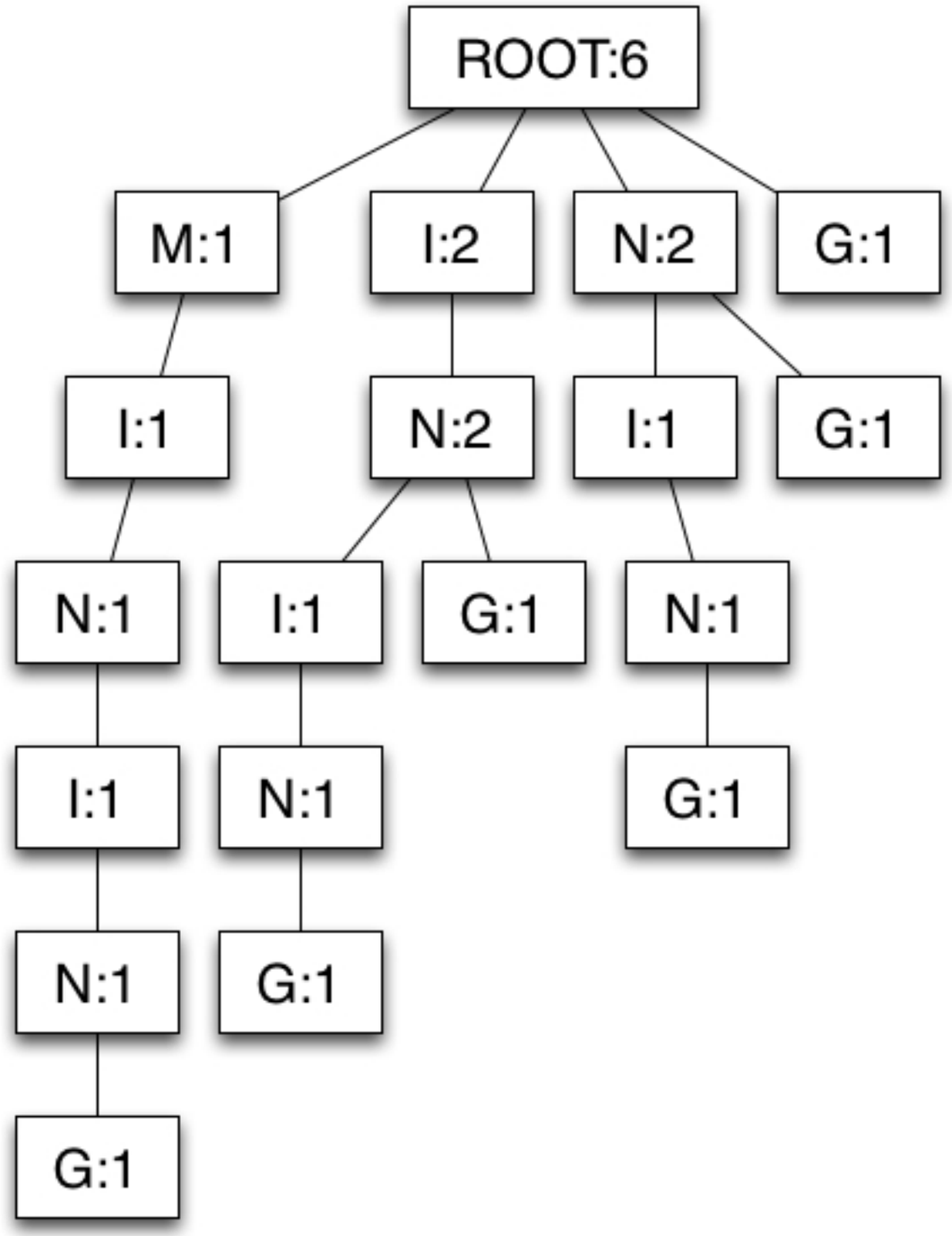
Аннотированное суффиксное дерево (Pampapathi, Mirkin, Levene, 2006):

- Средство представления частот фрагментов строки
- Строится за линейное время как любое СД ((Дубов, Черняк, 2013) – использование алгоритма Укконена для построения АСД)
- Занимает квадратичную от размера входа память



Свойства АСД

- Частота узла-родителя равна сумме частот узлов-детей
- Частота узла-родителя равна сумме частот листьев, которые он покрывает.



Основные меры релевантности строки тексту

- Релевантность в векторной модели
  - Косинусная мера близости

$$sim(s,d) = \frac{\sum_{i=1}^{n} s_i \times d_i}{\sum_{i=1}^{n} s_i^2 \sum_{i=1}^{n} d_i^2}$$

  - Okapi BM25

$$sim(s,d) = \sum_{i=1}^{n} IDF(s_i) \times \frac{f(s_i,d)(k_1+1)}{f(s_i,d) + k_1(1 - b + b\frac{|D|}{avgdl})}$$

$$IDF(s_i) = log\frac{N - n(s_i) + 0.5}{n(s_i) + 0.5}$$

- Релевантность в генеративных моделях (языковая модель и LDA)
  - Вероятность порождения строки моделью текста
- Релевантность в модели LSI $\hat{s} = \Sigma_k^{-1} U_{t \times k}^T s$

Релевантность строки тексту в модели аннотированного суффиксного дерева

- Мера релевантности (Pampapathi, Mirkin, Levene, 2006)
  - Построить АСД по тексту, а строку разбить строку на суффиксы
  - Оценка релевантности каждого суффикса

$$score(match(suffix, ast)) =$$

- $$= \sum_{node \in match} \varphi\left(\frac{f(node)}{f(parent(node))}\right)$$ ти

$$sim(s,d) = SCORE(s, ast) = \sum_{suffix} score(match(suffix, ast))$$

Шкалирующие функции

- линейная $\phi(x) = x$
- логистическая $\phi(x) = log(\frac{x}{1-x})$
- квадратный корень $\phi(x) = \sqrt{x}$

Нормированная оценка релевантности строки тексту в модели АСД (Миркин, Черняк, Чугунова, 2012)

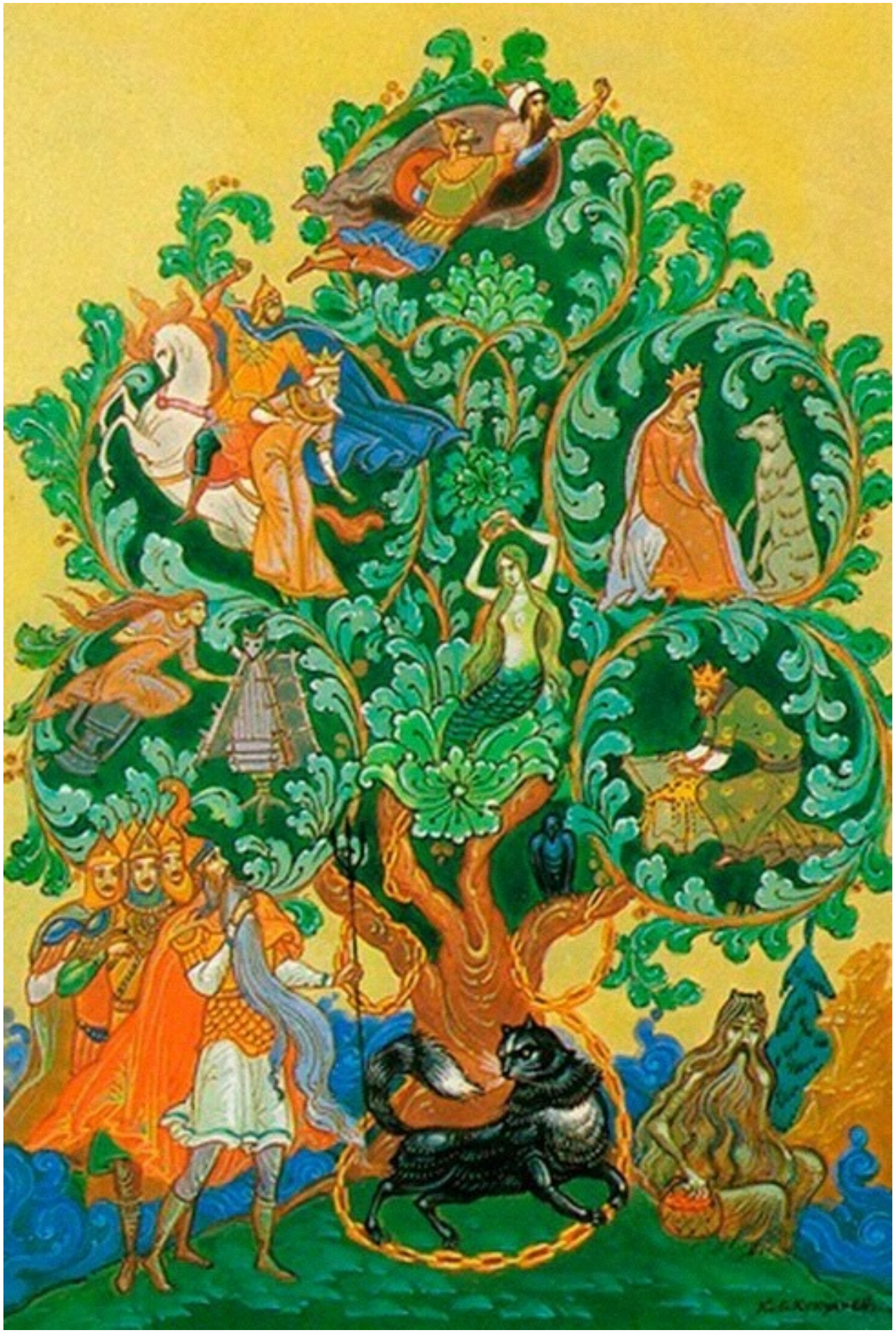
$$sim(s,d) = SCORE(s, ast) = \frac{\sum_{suffix} score(match(suffix, ast))/|suffix|}{|string|}$$

$$sim(dine, mining) = \frac{\left[{}^{\text{score(dine,}ast)}/_{4} + {}^{\text{score(ine,}ast)}/_{3} + {}^{\text{score(ne,}ast)}/_{2} + \text{score(e,}ast)\right]}{4} = 0 +$$

$$\frac{\varphi\left(\frac{2}{6}\right)+\varphi\left(\frac{2}{2}\right)}{3*4} + \frac{\varphi\left(\frac{2}{6}\right)}{2*4} + 0 = \frac{\frac{1}{3}+1}{12} + \frac{\frac{1}{3}}{8} = \frac{11}{72}.$$

Автоматизация производства

- **Автоматизация производства** — это процесс в развитии машинного **производства**, при котором функции управления и контроля, ранее выполнявшиеся человеком, передаются приборам и **автомати**ческим устройствам. Введение **автоматизаци**и на **производств**е позволяет значительно повысить **производ**ительность труда и качество выпускаемой продукции, сократить долю рабочих, занятых в различных сферах **производства**. До внедрения средств **автоматизаци**и замещение физического труда происходило посредством механизации основных и вспомогательных операций **производст**венного процесса. Интеллектуальный труд долгое время оставался не механизированным (ручным). В настоящее время операции физического и интеллектуального труда, поддающиеся формализации, становятся объектом механизации и **автоматизаци**и.



Аннотированное суффиксное дерево в задачах интерпретации текстов

- Задача: объяснение смысла текста входными строками
- Вход: совокупность строк и коллекция текстов
- Таблица строка–текст (pCT): по строкам – строки, по столбцам – тексты, в клетках – оценки релевантности

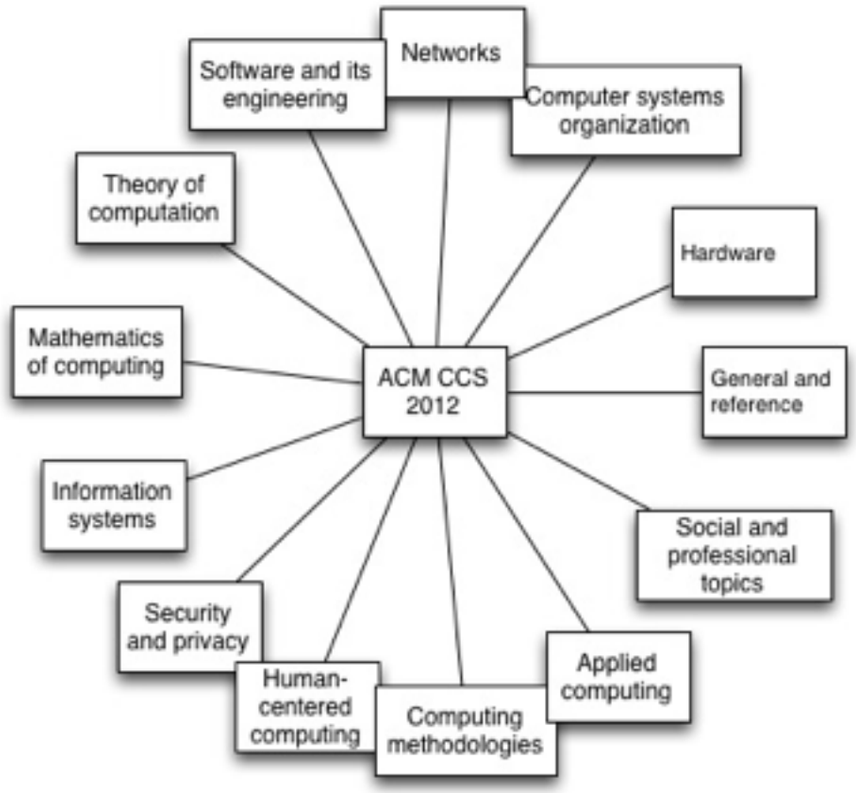
- Построение таблицы строка-текст:
  - Предварительная обработка: удаление мусора, каждый текст разбивается на 3-грамы
  - Для каждого текста строится свое АСД
  - Вычисляется релевантность каждой строки каждому АСД

Фрагмент pCT таблицы (Миркин, Черняк, Чугунова, 2010)

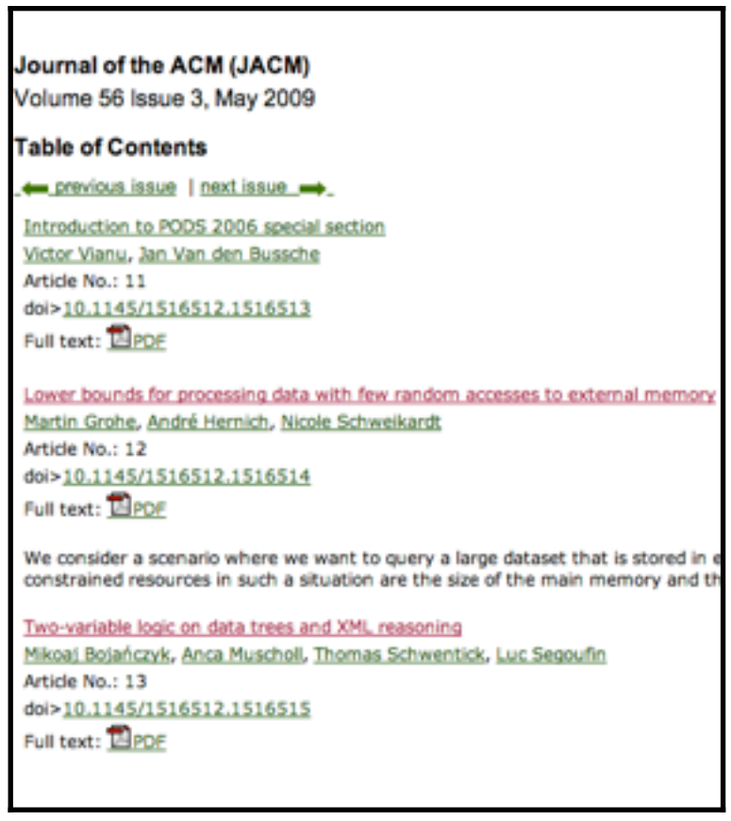
	Доклад Всеми	Междуна родные	Если генеральн
Изменение организацион	0.3145	0.3616	0.3644
Изменение уровня концентрации	0.5016	0.3148	0.2706
Повышение эффективнос	0.4433	0.2809	0.2445
Смена генерального	0.2264	0.2351	0.5947

Рубрикация научных статей

- **Вход**:
  - таксономия ACM-CCS 2012
  - коллекция аннотаций статей из журналов ACM (244 статьи из 3 журналов)



- **Построения:** pCT таблица таксономическая_тема X аннотация_статьи
- **Найти:** профиль каждой статьи
  - в профиль статьи включаем таксономические темы с высокими оценками



Journal of the ACM (JACM)
Volume 56 Issue 3, May 2009

Table of Contents

previous issue | next issue

Introduction to PODS 2006 special section
Victor Vianu, Jan Van den Bussche
Article No.: 11
doi>10.1145/1516512.1516513
Full text: PDF

Lower bounds for processing data with few random accesses to external memory
Martin Grohe, André Hernich, Nicole Schweikardt
Article No.: 12
doi>10.1145/1516512.1516514
Full text: PDF

We consider a scenario where we want to query a large dataset that is stored in e
constrained resources in such a situation are the size of the main memory and th

Two-variable logic on data trees and XML reasoning
Mikołaj Bojańczyk, Anca Muscholl, Thomas Schwentick, Luc Segoufin
Article No.: 13
doi>10.1145/1516512.1516515
Full text: PDF

Меры точности рубрикации статей

- Mean Average Precision (MAP)

$$AveP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{|relevant\_topics|}$$

$$MAP = \frac{\sum_{a \in abstracts} AveP(a)}{abstracts},$$

- normalised Discounted Cumulative Gain (nDCG)

$$nDCG_k = \frac{DCG_k}{IDCG_k}, \text{ где}$$

$$IDCG_k = rel(1) + \sum_{i=2}^{|relevant\_topics|} \frac{1}{log_2 i}$$

$$DCG_k = rel(1) + \sum_{i=2}^{k} \frac{rel(i)}{log_2 i}$$

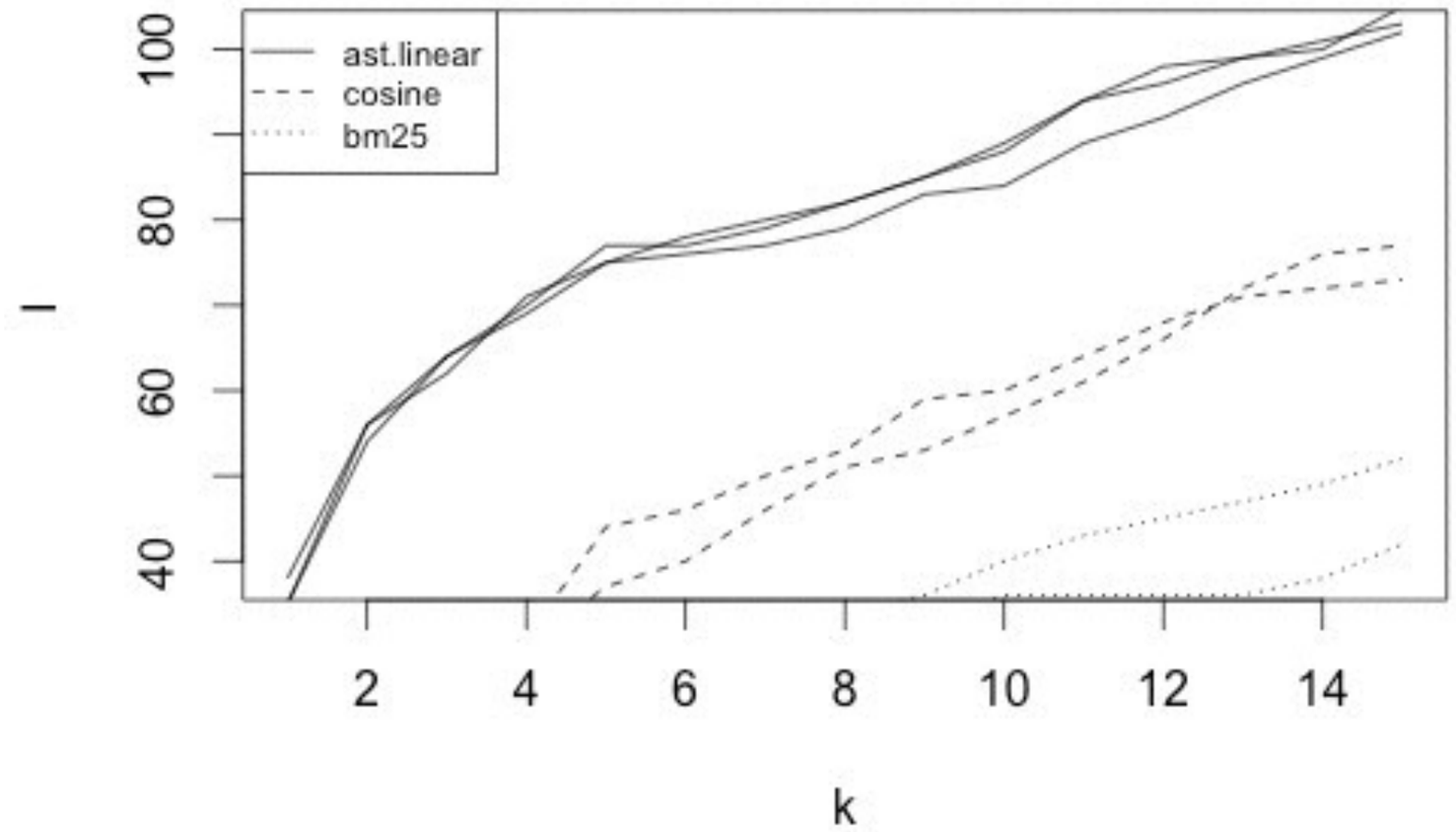
- Intersection at K

$$I_k = \sum_{a \in abstracts} i(a)$$

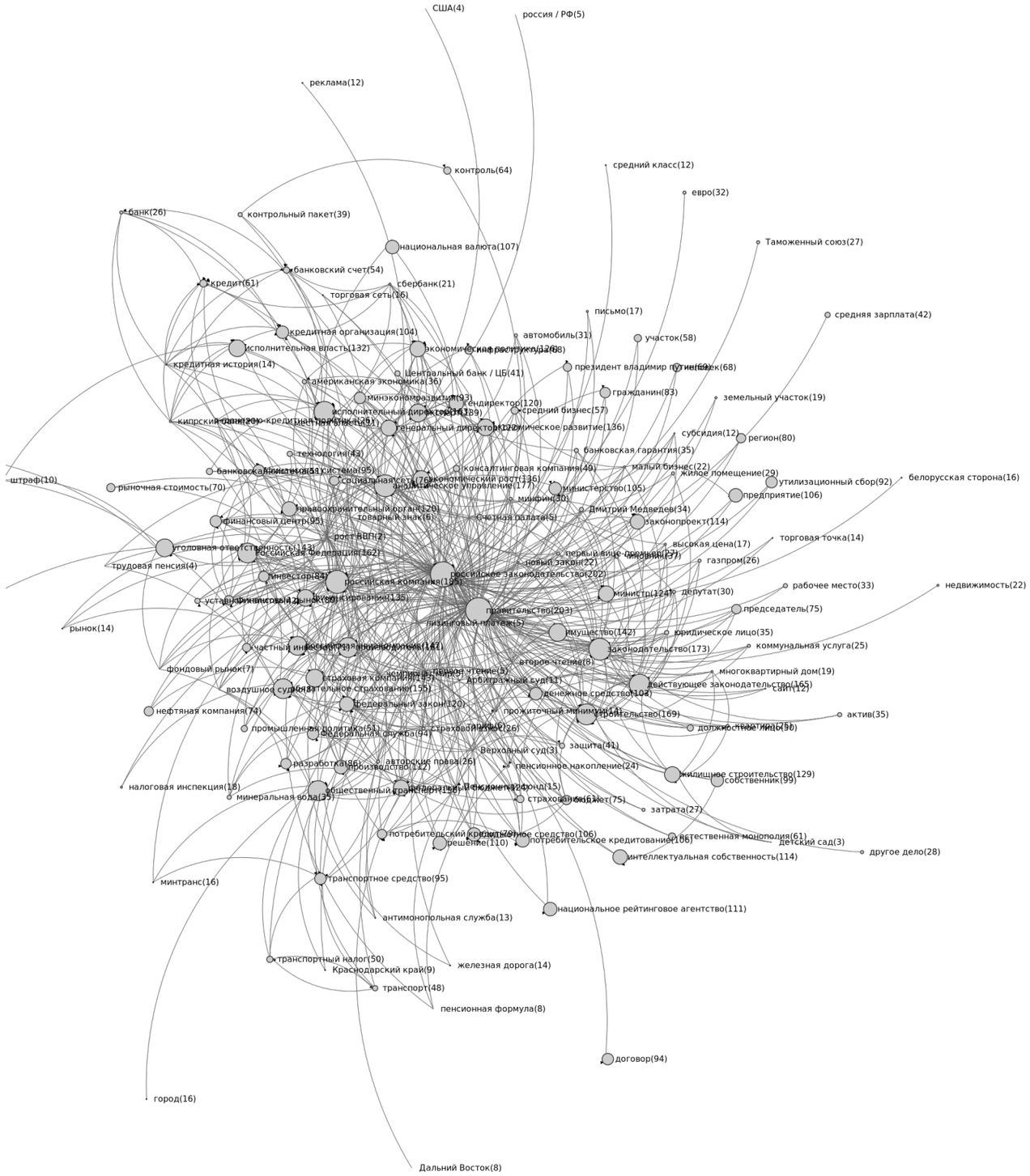
Результаты – 1

Preprocessing technique	I_k		MAP_{15}	$nDCG_{15}$
	I_5	I_{15}		
Cosine relevance measure				
words	44	73	0.0748	0.0245
stems	37	77	0.0788	0.0250
coll3	41	76	0.0911	0.0278
coll3.5	31	71	0.0642	0.0237
BM25 relevance measure				
words	14	52	0.0631	0.0279
stems	21	36	0.0869	0.0259
coll3	15	46	0.0524	0.0224
coll3.5	16	46	0.0577	0.0228
CPAMF relevance measure				
linear.0	75	102	0.3588	0.1124
linear.1	77	105	0.3550	0.1133
linear.2	75	103	0.3486	0.1120
root.0	75	102	0.3657	0.1125
root.1	77	104	0.3561	0.1122
root.2	77	106	0.3497	0.1126
logit.0	36	57	0.1214	0.0450
logit.1	18	33	0.0521	0.0216
logit.2	29	56	0.0780	0.0335

Результаты – 2



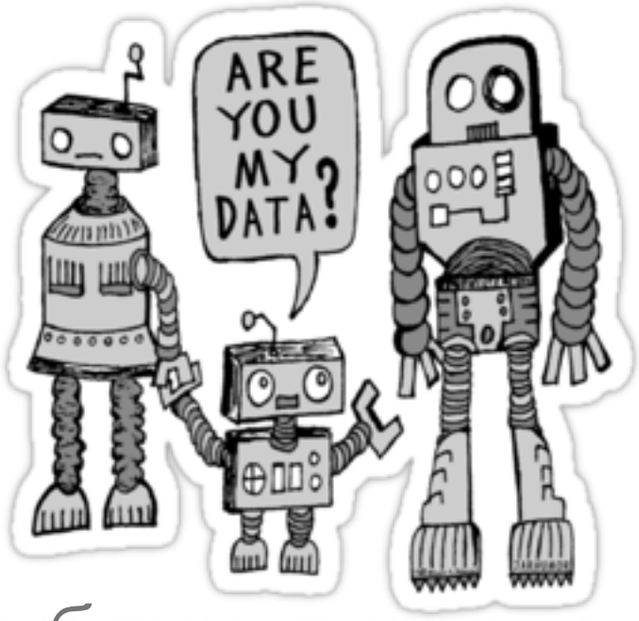
Построение концептуальных карт



2 концептуальные карты “Бизнес”

- Ключевые словосочетания заданы экспертами

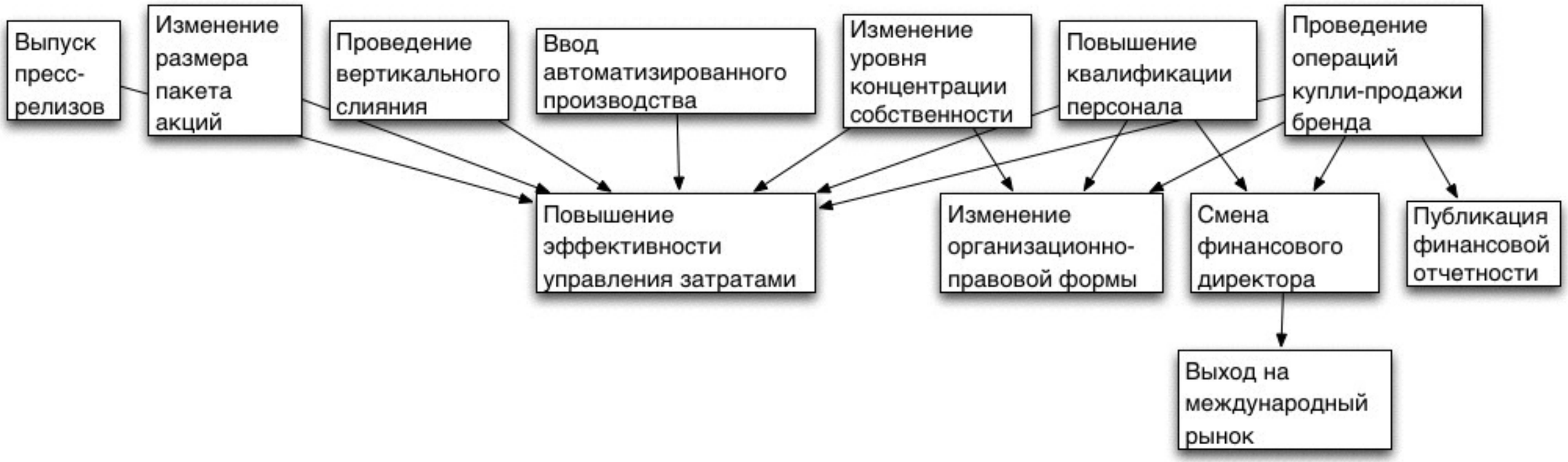
Раскрытие ошибок отчетности
Смена генерального директора
Участие в судебных разбирательствах
Выход на IPO
Продажа активов
Сужение бизнеса



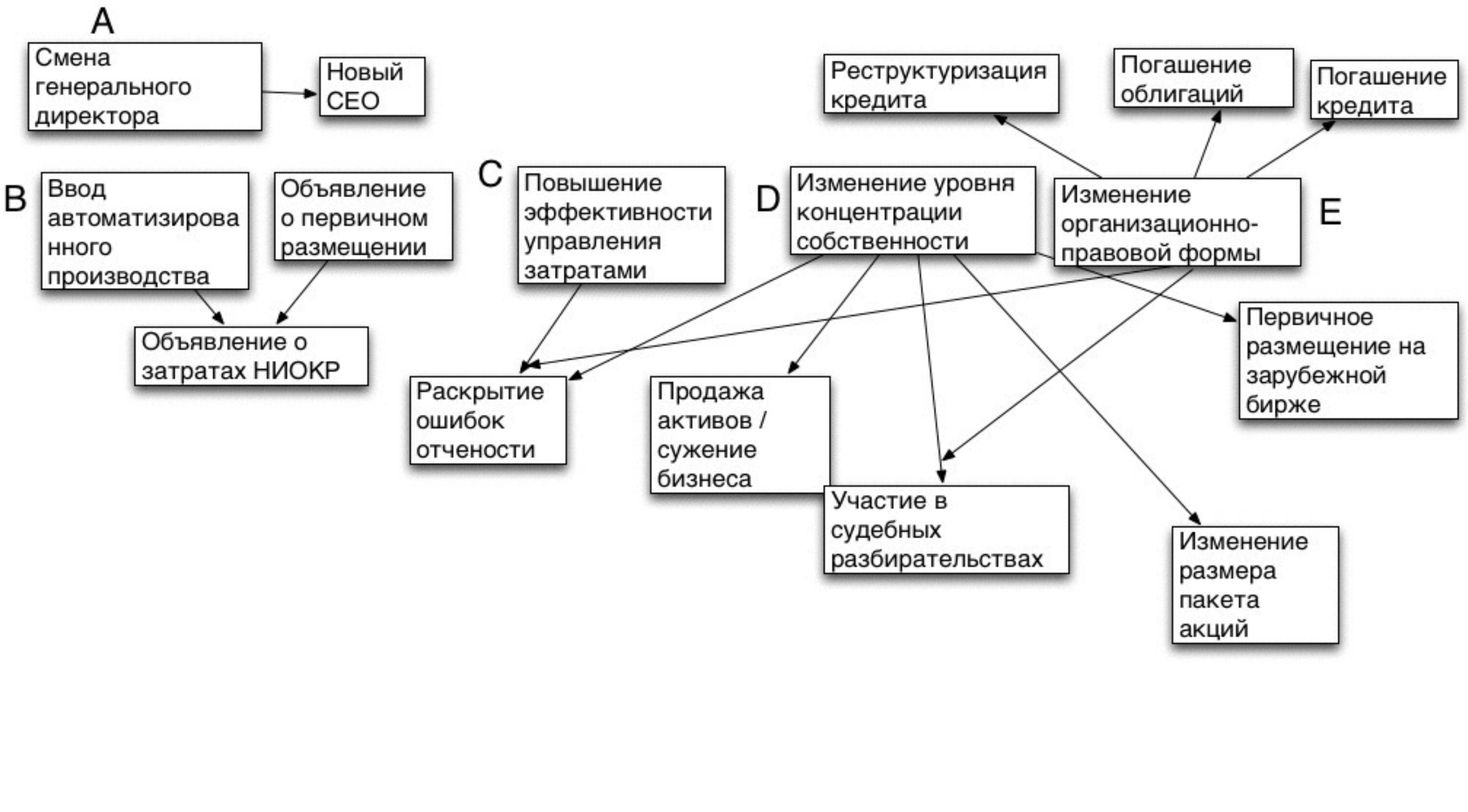
- Коллекция новостных сообщений, опубликованных на порталах газет (~1000 текстов)

- Статьи из Википедии, принадлежащие к категории “Бизнес” (~12000 текстов)

Газетная концептуальная карта



Энциклопедическая концептуальная карта



Метод построения концептуальной карты – 1

- Оцениваем релевантность каждого ключевого словосочетания каждому тексту

- Определяем для каждого ключевого словосочетания K_i множество текстов $F(K_i)$, которым оно релевантно

- Порог релевантности: r (из интервала $[0.5,1)$)

Метод построения концептуальной карты – 2

- $K_i \implies K_j$, если

$$\frac{|F(K_i) \cap F(K_j)|}{|F(K_i)|} > c$$

- c – порог на поддержу: $c \in [0.5, 1)$

- **Пример**: Ввод автоматизированного производства ⇒ Повышение эффективности управления затратами

Дальнейшее развитие

- Автоматизация выделения ключевых словосочетаний
- Использование методов теории графов для анализа концептуальных карт
- Расширение на случай правил вида многие-к-одному, многие-ко-многим, один-к-многим

Достраивание таксономии на основе ресурсов Википедии

- **Вход**:
  - фрагмент таксономии MIAMI, построенной вручную
  - фрагмент дерева категорий Википедии
  - коллекция статей Википедии

- **Построения:** четыре pCT таблицы название_статьи X статья, название_категории X статья, название_родительской_категории X статья, таксономическая_тема X категория

- **Очистить** дерево категорий от иррелевантных статей и категорий

- **Достроить** промежуточные уровни таксономии

Фрагмент таксономии MIA по материалам паспортов ВАК

ТВиМС	Теория вероятностей и математическая статистика		
	ТВиМС.01	Теория вероятностей	
		ТВиМС.01.01	Модели и характеристики случайных явлений
		ТВиМС.01.02	Распределения вероятностей и предельные теоремы
		ТВиМС.01.03	Комбинаторные и геометрические вероятностные задачи
		ТВиМС.01.04	Случайные процессы и поля
		ТВиМС.01.05	Оптимизационные и алгоритмические вероятностные задачи
	ТВиМС.02	Математическая статистика	
		ТВиМС.02.01	Методы статистического анализа и вывода
		ТВиМС.02.02	Статистические параметры и их оценивание по выборке
		ТВиМС.02.03	Статистические критерии и проверка статистических гипотез
		ТВиМС.02.04	Временные ряды и случайные процессы
		ТВиМС.02.05	Машинное обучение
		ТВиМС.02.06	Многомерная статистика и анализ данных

Фрагмент дерева категорий

Математическая статистика		
	Факторный анализ	
		Коэффициент детерминации
		Метод главных компонент
		Линейная регрессия на корреляции
		Факторный анализ
		Коррелятор
		RANSAC
		Метод максимального правдоподобия
		Метод группового учета аргументов
		Мультиколлинеарность
		Метод моментов нахождения оценок
		Робастность в статистике
		Корреляция

Статья Википедии

Факторный анализ

Материал из Википедии — свободной энциклопедии

Текущая версия страницы пока не проверялась опытными

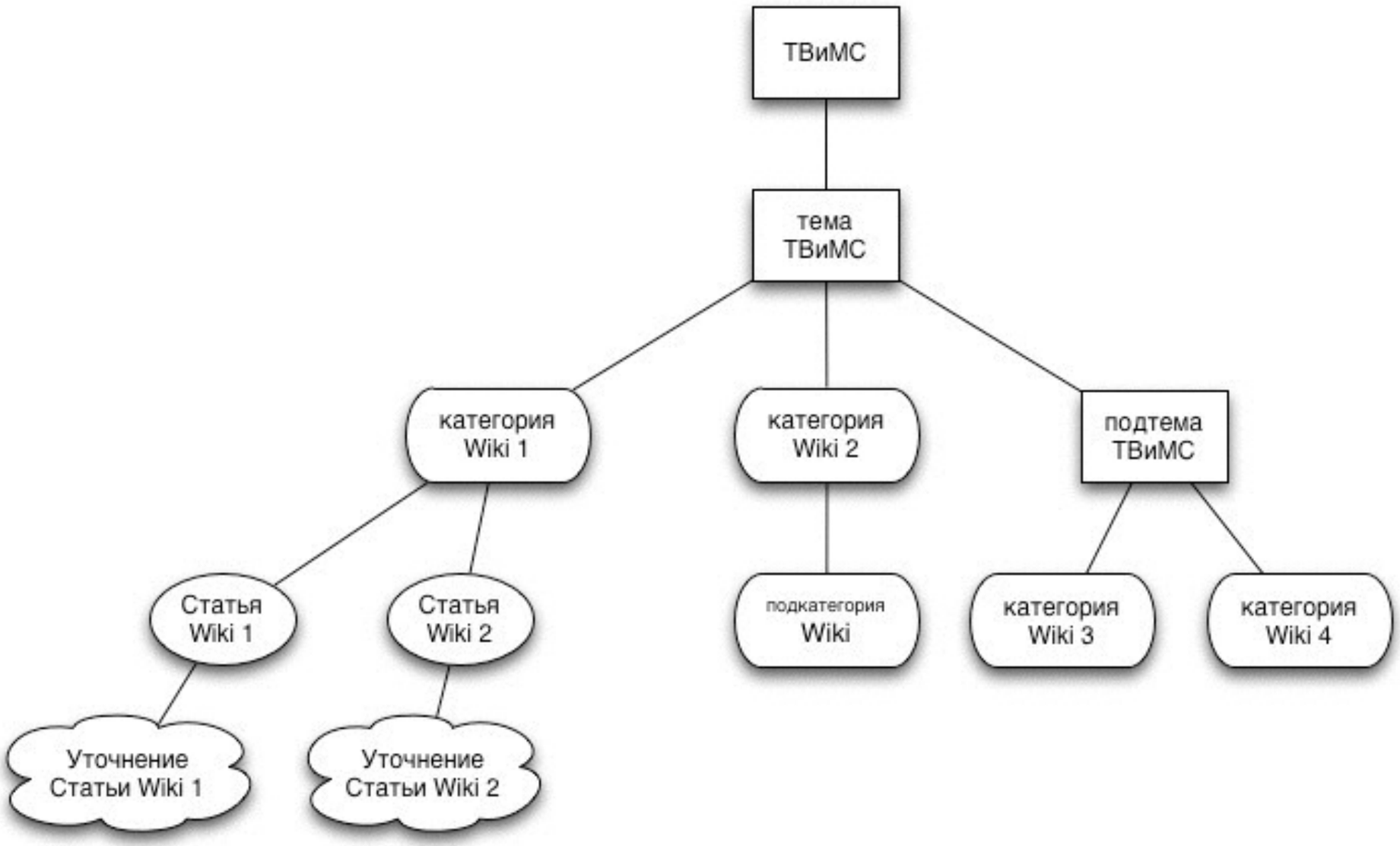
Факторный анализ — многомерный метод, применяемый для изучен
количества неизвестных переменных и случайной ошибки.

Содержание [убрать]

Краткая история

Факторный анализ впервые возник в психометрике и в настоящее вре
статистике и других науках. Основные идеи факторного анализа был
большой вклад в исследование индивидуальных различий. Но в разра
занимались такие ученые как Спирмен Ч. (1904, 1927, 1946), Терстоун
Пирсона К., в значительной степени развившего идеи Ф. Гальтона, ам
заслуживает и английский психолог Айзенк Г., широко использовавш
разрабатывался Хотеллингом, Харманом, Кайзером, Терстоуном, Так
Statistica и т. д.

Предлагаемая структура достраиваемой таксономии (по аналогии с ACM-CCS)



Этапы достраивания таксономии

- Извлечение фрагмента дерева категорий Википедии
- Подготовка машинного представления дерева категорий и таксономии
- Предварительная подготовка текстов статей: превращение каждой в последовательность строк
- Очистка дерева категорий от иррелевантных статей
- Очистка дерева категорий от иррелевантных категорий
- Достраивание промежуточных уровней таксономии
- Извлечение уточняющих слов и словосочетаний из текстов статей по шаблонам (СУЩ, ПРИЛ + СУЩ)

Использование метода АСД в задаче достраивания таксономии

1.Для очистки дерева категорий от иррелевантных статей:

- релевантность названия статьи тексту статьи
- релевантность названия родительской категории тексту статьи

2.Для очистки дерева категорий от иррелевантных категорий:

- релевантность названия категориив совокупности статей данной категории

3.Для достраивания категорий к промежуточным уровням таксономии:

- релевантность таксономической темы совокупности статей категории (категория достраивается к таксономической теме наиболее ей релевантной)
- релевантность названия категории совокупности статей подкатегории (если подкатегории более релевантно название родительской категории, оставляем ее на промежуточных уровнях)

Очистка дерева категорий от иррелевантных статей

Статьи категории Факторный анализ	
факторный анализ	0.561
метод максимального правдоподобия	0.529
корреляция	0.349
коэффициент детерминации	0.231
метод главных компонент	0.207
линейная регрессия на корреляции	0.157
коррелятор	0.143
RANSAC	0.097
Робастность в статистике	0.067

Очистка дерева категорий от иррелевантных категорий

Подкатегориии категории Машинное обучение	
методы обучения нейросетей	0.278
деревья принятия решений	0.180

Достраивание категорий к промежуточным уровням таксономии

ТВиМС.01.02	распределения вероятностей и предельные теоремы		
	средние величины		0.429
	распределения вероятностей		0.445
		дискретные распределения	
		непрерывные распределения	
	марковские процессы		0.474
	мартингалы		0.476

ТВиМС

ТВиМС.02
Математическая
статистика

Факторный
анализ

Коэффициент
детерминации

Метод главых
компонент

факторный
анализ

Метод
максимального
правдоподобия

Корреляция

модель
коэффициент
детерминации
мнк
объясняющая
переменная
фактор
штраф
лаг
метод
наименьших
квадратов

главная
компонета
проекция
дисперсия
размерность
координата
нормировка
апроксимация
данных
тензор
сингулярное
разложение

факторный
анализ
фактор
нагрузка

максимальное
правдоподобие
функция
правдоподобия
параметр
выборка
оценка
максимального
правдоподобия

корреляция
коэффициент
корреляции
случайная
величина
ковариация
связь
ранг
отклонение

Заключение – 1

- Достоинства АСД меры релевантности:
  - учет нечетких совпадений
  - нормированные оценки релевантности строки-тексту, независимые от количества слов в тексте
- Недостатки:
  - Вычислительная сложность
  - Отсутствие мер качества результатов

Заключение – 2

- Другие задачи, где можно использовать АСД:
  - вычисление близости между предложениями в задаче суммаризации текстов
  - порождение n-грам для классификации по жанрам
  - разделение сложных слов на составляющие
  - сравнение с string kernel

ГРНТИ. (2014, 08 13).
Дубов, М. С., & Черняк, Е. (2013). Аннотированные суффиксные деревья: особенности реализации . *АИСТ* (pp. 10-20). Екатеринбург: Национальный Открытый Университет «ИНТУИТ» .
Агеев, М. С., Добров, Б. В., & Лукашевич, Н. В. (2008). Автоматическая рубрикация текстов: методы и проблемы. *Ученые записки Казанского государственного университета, серия Физико-математические науки , 150* (4), 25-40.
Байтин, А. (2008). Испарввление поисковых запросов в Яндексе. Вероятностная языковая модель. *Российские Интернет-Технологии.*
Миркин, Б., Черняк, Е., & Чугунова, О. (2012). Метод аннотированного суффиксного дерева для оценки степени вхождения строк в текстовые документы. *Бизнес–информатика , 21* (3), 31-41.
Паспорта научных специальностей ВАК. (n.d.). Retrieved from http://old.mon.gov.ru/work/nti/dok/vak/11.11.11-pasporta.pdf
Воронцов, К. В. (n.d.). Лекции по вероятностностному тематическому моделированию.
Лукашевич, Н. В. (2011). *Тезаурусы в задачах информационного поиска.* Издательство МГУ.
Сегалович, И. (2002). Как работают поисковые системы? *Мир Интернет* (10).
Классификатор ВАК. (n.d.). Retrieved from http://www.viniti.ru/russian/math/files/271.htm
Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *the 1993 ACM SIGMOD international conference on Management of data.* 207-217.
Andrews, N., & Fox, E. (2007). *Recent Developments in Document Clustering.* Retrieved 10 10, 2014, from http://www.hse.ru/data/2012/12/20/1303727883/2.pdf
Arora, R., & Ravindran, B. Latent dirichlet allocation based multi-document summarization. *Proceedings of the second workshop on Analytics for noisy unstructured text data*, (pp. 91-97). 2008.
Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, (pp. 722-735). 2007.
Berry, M. W., & Browne, M. (2005). *Understanding Search Engines: Mathematical Modeling and Text Retrieval.* Philadelphia: SIAM.
Berry, M. W., Dumais, S., & O'Brien, G. (1995). Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review , 37*, 573-595.
Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* Sebastopol: O'Reilly Media Inc.
Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media .
Bíró, I., Szabó, J., & Benczúr, A. (2008). Latent dirichlet allocation in web spam filtering. *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, (pp. 29-32).
Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research , 3* (4-5), 993–1022.
Cantador, I., Bellogin, A., & Vallet, D. (2010). Content-based recommendation in social tagging systems. *4th ACM conference on Recommender systems*, (pp. 237–240). Barcelona.
CCS. (2014, 06 09). *http://www.acm.org/about /class/2012* .
Ceci, M., & Malerba, D. (2007). Classifying web documents in a hierarchy of categories: a comprehensive study. *Journal of Intelligent Information Systems , 28* (1), 37-78.
Chernyak, E. L., & Mirkin, B. G. A method for refining a taxonomy by using annotated suffix trees and Wikipedia resources. *2nd International Conference on Information Technology and Quantitative Management,* (pp. 193–200). Moscow: Elsevier.
Cui, C., Lu, Q., Li, W., & Chen, Y. (2009). Mining concepts from Wikipedia for ontology construction. *the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (pp. 287-290). Washington, DC, USA: IEEE Computer Society,.
Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. *SAC '03 Proceedings of the 2003 ACM symposium on Applied computing* (pp. 784-788). New York, NY, USA: New York, NY, USA.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the american society for information science , 4* (6), 391-407.
Ding, C. (2010). *A survey of ontology construction and information extraction from Wikipedia.* University of Southampton .
Duh, K., & Kirchhoff, K. (2008). Learning to rank with partially-labeled data. *31st Annual International ACM SIGIR Conference*, (pp. 251-258). Singapore.
Gee, R. K. (2004). Using latent semantic indexing to filter spam. *Proceedings of the 2003 ACM symposium on Applied computing*, (pp. 460-464).
Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19-25). ACM.
Gupta, A., & Kumaraguru, P. (2012). Credibility Ranking of Tweets during High Impact Events. *1st Workshop on Privacy and Security in Online Social Media*, (pp. 2-8). Lyon.
Gusfield, D. (1997). *Algorithms on strings, trees and sequences: computer science and computational biology.* Cambridge: Gambridge University Press.
Harris, Z. (1954). Distributional structure . *Word , 10* (23), 146–162.
Harris, Z. (1954). Distributional Structure. *Word , 10* (23), 146-162.
Hjørland, B. (2010). The founderation of the concept of relevance. *Journal of the American Society for Information Science and Technology , 61* (1), 217-237.
Hoffart, J., Suchanek, F., Berberich, K., & Weikum, G. (2013). YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* , 28-61.
Hoffman, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems , 22* (1), 89-115 .
Hwang, G.-J. (2003). A conceptual map model for developing intelligent tutoring systems. *Computers & Education , 4* (3), 217-235.
Hyunsoo, K., Howland, P., & Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research* , 37-53.
Ifenthaler, D. (2010). Relational, structural, and semantica nalysis of graphical representations and concept maps. *Educational Technology Research and Development, , 58* (1), 88-97.
Jiang, S., Bing, L., & Zhang, Y. (2013). Towards an enhanced and adaptable ontology by distilling and assembling online encyclopedias. *the 22nd ACM International Conference on Information & Knowledge Management.*, (pp. 1703-1708).
Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing.* Prentice Hall.
Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing , 35* (3).
Kittur, A., Chi, E., & Suh, B. (2009). What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. *the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 1509-1512). Boston, USA.
Koehn, P., Och, F., & Marcu, D. (2003). Statistical phrase based translation. *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics.*
Krestel, R., Frankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. *Proceedings of the third ACM conference on Recommender systems* (pp. 61-68). ACM.
Liu, X., Song, Y., Liu, S., & Wang, H. Automatict axonomy construction from keywords. *the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1433-1441).

Yu, C., & Salton, G. (1976). Precision Weighting - An Effective Automatic Indexing Method. *Journal of ACM , 23* (1), 76-88.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.
Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.
Mirkin, B. G. (2011). *Core Concepts in Data Analysis: Summarization, Correlation and Visualization.* Heidelberg: Springer.
Mishra, M., Huan, J., Bleik, S., & Song, M. (2012). Biomedical text categorization with concept graph representations using a controlled vocabulary. *11th International Workshop on Data Mining in Bioinformatics* (pp. 26-32). New York, NY,: ACM.
Monay, F., & Gatica-Perez, D. (2003). On image auto-annotation with latent space models. *Proceedings of the eleventh ACM international conference on Multimedia.*
Pampapathi, R., Mirkin, B., & Levene, M. (2006). A suffix tree approach to anti-spam email filtering. *Machine learning , 65* (1), 309-338.
Pang, B., Lee, L., & Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* , (pp. 79–86). 2002.
Pantel, P., & Lin, D. (2002). Discovering word senses from text . *roceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* , (pp. 613–619). Edmonton, Canada .
Pissanetzky, S. (1984). *Sparse Matrix Technology.* Academic Press.
Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. *ACM SIGIR conference on Research and development in information retrieval* (pp. 275-281). New York, NY, USA: ACM.
Ponzetto, S., & Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. *In proceedings of AAAI conference on artificial intelligence*, (pp. 78-85). Vancouver, Canada.
Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems. , 14* (3), 130–137.
Raghavan, V. V., & Won, S. K. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science , 37* (5), 279 - 287.
Ramage, D., Hall, D., Nallapati, R., & Manning, C. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 1*, pp. 248-256. Association for Computational Linguistics.
Ranwez, S., Ranwez, V., Villerd, J., & Crampes, M. (2006). Ontological distance measures for information visualisation on conceptual maps. *1st International conference on the Move to Meaningful Internet Systems*, (pp. 1050-1061). Heidelberg.
Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity . *Proceedings of the Ninth Machine Translation Summit* , (pp. 315–322).
Reed, J. W., Jiao, Y., Potok, T. E., Klump, B. A., Elmore, M. T., & Hurson, A. R. (2006). TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams. *ICMLA '06 Proceedings of the 5th International Conference on Machine Learning and Applications* (pp. 258-263). Washington, DC, USA: IEEE Computer Society.
Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (pp. 45-50).
Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. . *Foundations and Trends in Information Retrieval , 3* (4), 333-389.
Robinson, P., & Bauer, S. (2011). *Introduction to bio-ontologies.* USA: Chapman & Hall / CRC .
Sadikov, E., Madhavan, J., Wang, L., & Halevy, A. (2008). Clustering query refinements by user intent. *19th International Conference on World Wide Web*, (pp. 841-850). New York, USA.
Salton, G., & Buckley, C. (1998). Term-weighting approaches in automatic text retrieval. *nformation Processing and Management , 25* (5), 513 – 523.
Santos, A., & Rodrigues, F. (2010). Multi-Label Hierarchical Text Classification Using the ACM Taxonomy . *14th Portuguese Conference on Artificial Intelligence* , (pp. 553 – 564). Aveiro.
Sebastiani, F. (2002). Machine learning in automated text categorization. *Journal of ACM Computing Surveys , 34* (1), 1-42.
Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Pacific Grove, CA: Brooks Cole Publishing Co.
Strapparava, C., & Valitutti, A. WordNet-Affect: An affective extension of WordNet. *Proceedings of LREC*, (pp. 1083–1086).
Taylor, A. (2012). User relevance criteria choices and the information search process. *Information Processing & Management* , 135-153.
Tseng, S., Sue, P., Su, J., Weng, J., & Tsai, W. (2007). A new approach for constructing the concept map. *Computers & Education , 49* (3), 691-707.
Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the Association for Computational Linguistics*, (pp. 417–424).
Turney, P. D., Littman, M. L., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. *Proceedings of the International Conference on Recent Advances in Natural Language Processing* , (pp. 482–489). Borovets, Bulgaria .
Turney, P., & Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research , 37* (1), 141-188.
Valizadegan, H., Jin, R., Zhang, R., & Mao, J. (2010). Learning to Rank by Optimizing NDCG Measure. *Advances in Neural Information Processing Systems* , 1883-1891.
Van Hage, W., Katrenko, S., & Schreiber, G. (2005). A method to combine linguistic ontology-mapping techniques. *the 4th International Semantic Web Conference*, (pp. 34-39). Galway, Ireland.
Wang, X., & Grimson, E. (2008). Spatial latent dirichlet allocation. *Advances in Neural Information Processing Systems* .
Wei, X., & Croft, B. (2006). LDA-based document models for ad-hoc retrieval. *Proceeding SIGIR '06 Proceedings of the 29th annual international ACM* (pp. 178 - 185). New York, USA: ACM.
White, R., Bennett, P., & Dumais, S. (2010). Predicting short-term interests using activity-based search contexts. *19th ACM Conference on Information and Knowledge Management*, (pp. 1009-1018). Toronto, Canada.
Wong, S. K., Ziarko, W., & Wong, P. C. (1985). Generalized vector spaces model in information retrieval. *the 8th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 18-25). New York, USA: ACM.
Xavier, C. C., & De Lima, V. L. (2011). A semi-automatic method for domain ontology extraction from portuguese language Wikipedia's categories. *Advances in Artificial Intelligence* , 11-20.
Xia, F., Liu, T., Wang, J., Zhang, W., & Li, H. (2008). Listwise approach to learning to rank - theory and algorithm. *25th International Conference on Machine Learning*, (pp. 1192–1199). Helsinki.