Scoring string-to-text relevance: method and applications

Ekaterina Chernyak and Boris Mirkin
Division of Applied Mathematics and Informatics
NRU HSE, Moscow, Russia

Motivation

- Некоторые распространенные задачи:
  - Аннотирование статьи таксономическими темами
  - Построение таксономии по коллекции текстов (например, статьи Википедии)

- Новый метод аггрегирования коллекции текстов: представление текста ключевыми слловосочетаниями, а не таксономией / таксономическими темами

- Следовательно, возникает потребность в мере релеватности с:
  - минимальной предварительной подготовкой текстов (суффиксные деревья)
  - независящей от различий в длинах текстов или строк

Аннотирование статьи таксономическими темами

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

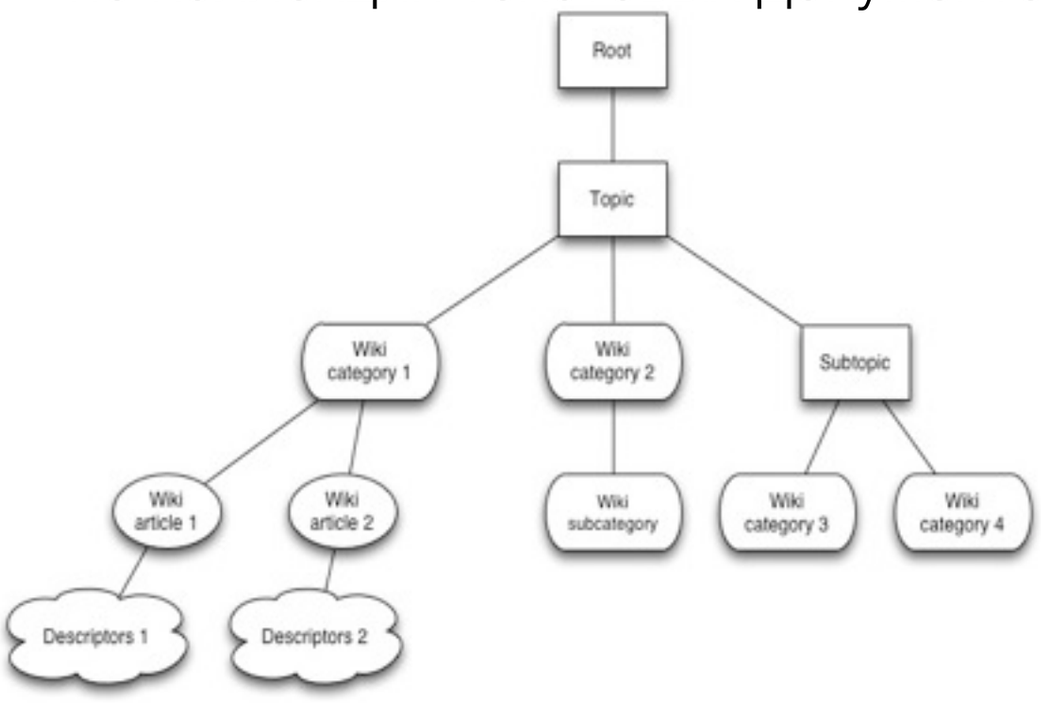
Two-variable logic on data trees and XML reasoning
Mikoaj Bojańczyk, Anca Muscholl, Thomas Schwentick, Luc Segoufin
Article No.: 13
doi>10.1145/1516512.1516515
Full text: PDF

В этих трех задачах возникает необходимость...

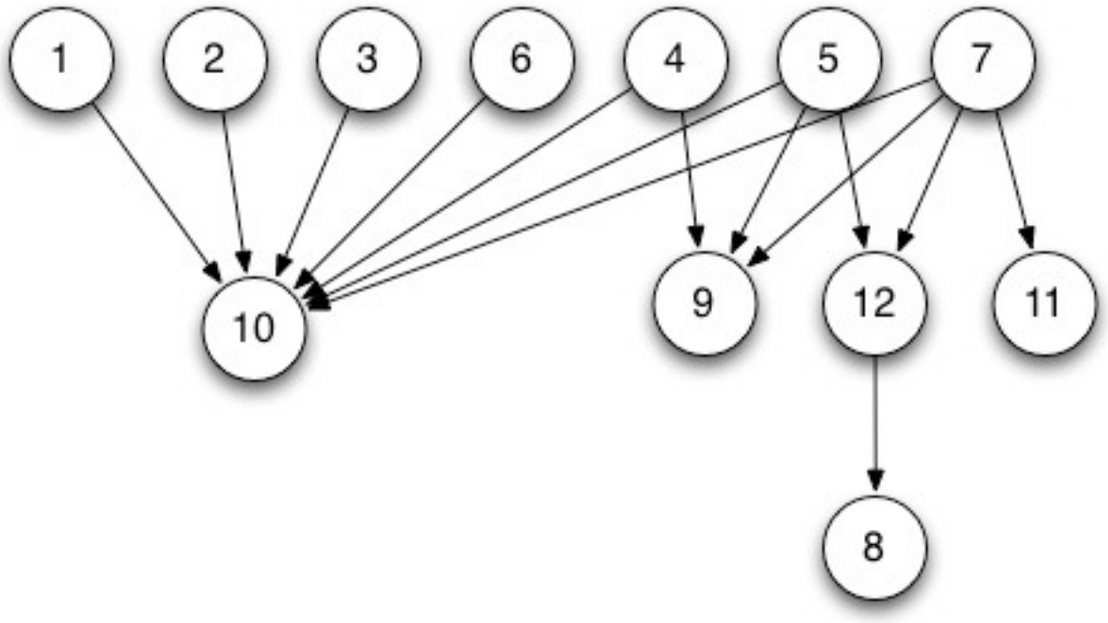
Primary Classification: F.1.1
Additional Classification: F.1.3, H.2.4

Primary Classification: F.4.1
Additional Classification: F.4.3, H.2.1, H.2.3, I.7.2

Построение таксономии по коллекции текстовых документов



Построение графа связей между ключевыми словосочетаниями



... В мере релеватности строки тексту

Строка \ Текст	Доклад Всемирного Банка об экономике России	Международные стандарты финансовой отчетности	Если генеральный директор иностранец
Изменение организационно-правовой формы	0.3145	0.3616	0.3644
Изменение уровня концентрации собственности	0.5016	0.3148	0.2706
Повышение эффективности управления затратами	0.4433	0.2809	0.2445
Смена генерального директора	0.2264	0.2351	0.5947

Используем СТ таблицу (строка Х текст)

Цели

- Разработать метод для оценивания релевантность ключевогого словосочетания неструктурированному тексту и
- Применить его к представленным выше задачам

Входные данные

- Коллекция текстовых документов
- Множество ключевых слов и словосочетания
  - В т.ч. таксономия - иерархия ключевых слов и словосочетаний

Основные этапы

- Предварительная подготовка текстов
- Использование метода аннотированного суффиксного дерева (АСД) для построения CT таблицы
- Анализ CT таблицы

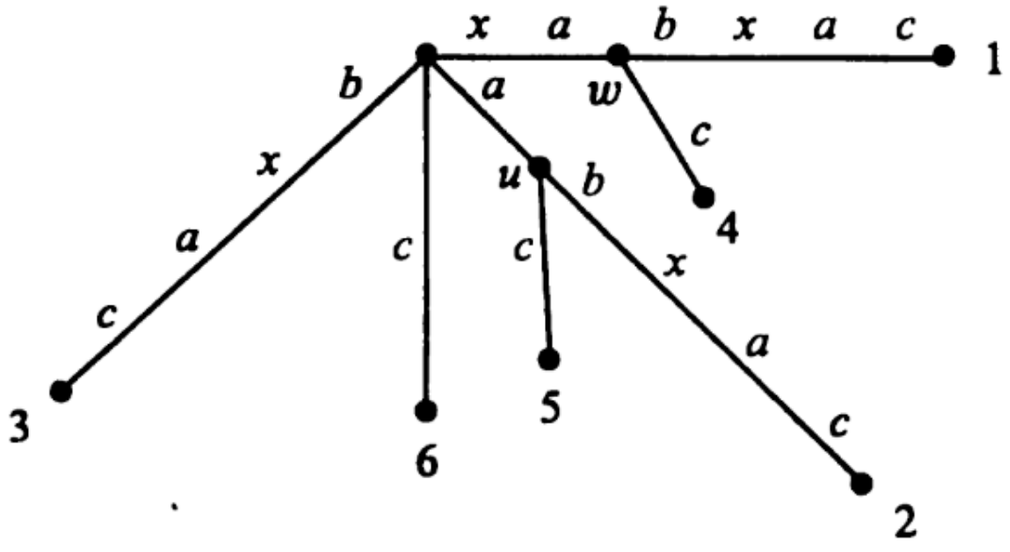
Текст –> множество коротких строк

- Разбиваем текст на строки
    - Токенизация
    - Построение частотных словарей n-грам (n=2,3...)
- Нет потребности в стэминге, морфологическом и синатксическом анализе!

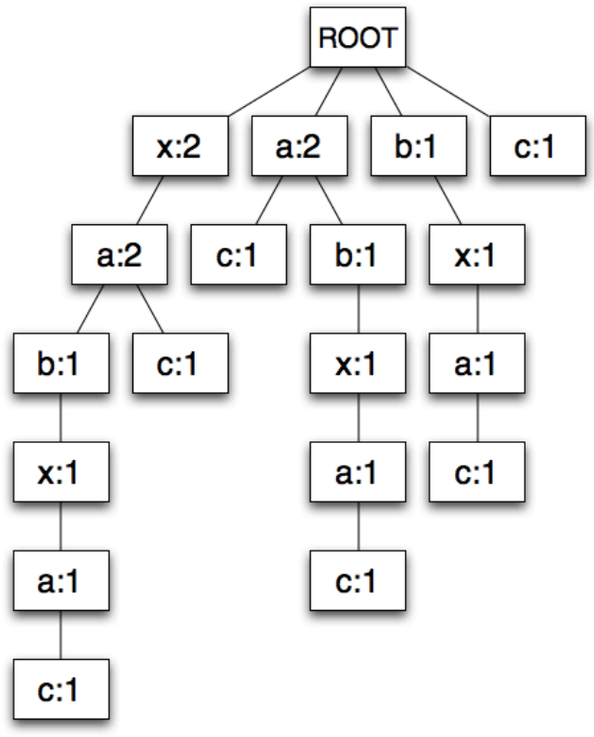
Суффиксное дерево

Суффиксное дерево (Weiner, 1973): средство хранения строковых данных	Аннотированное суффиксное дерево (Pampapathi, Mirkin, Levene, 2006): средство представления частот фрагментов строки
•**Ребра** помечены суффиксами строки •У каждого узла два или больше потомка •Листья пронумерованы	•Каждый узел помечен **одним** символом •и аннотирован его частотой

‘xabxac’

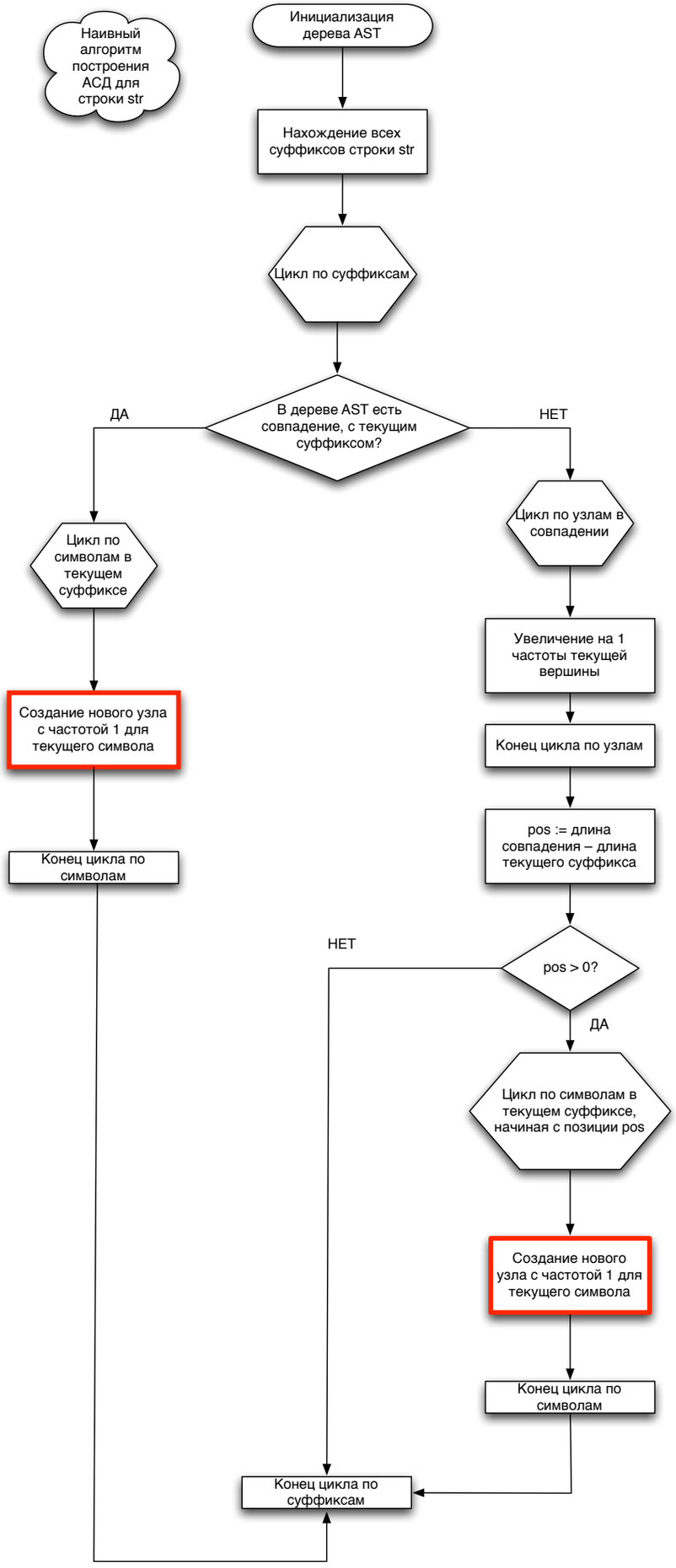


‘xabxac’



Построение АСД

- Найти все суффиксы данной строки
- Построить цепочку узлов для первого суффикса
- Для последующих суффиксов:
  - найти совпадение, увеличить частоты в совпадении, создаем узлы, если совпадение короче суффикса
  - если совпадение не найдено, создать новую цепочку узлов



Оценка сходства строки множеству строк

- ast – АСД с узлами u , построенное по множеству строк $coll$
- s – строка длины l
- $f(u)$ – частота узла u , $\sum_{i \in n_u} f(i)$ – частота отца u
- Условная вероятность u – $\hat{p}(u) = \frac{f(u)}{\sum_{i \in n_u} f(i)}$
- $m = m_0 m_1 .... m_k$ – совпадающие узлы, k – число узлов в совпадении
- Оценка совпадения – $score(m) = \frac{\sum_{i=0}^{k} \hat{p}(m_i)}{k}$
- Окончательная оценка – $SCORE(s, ast) = \frac{\sum_{i=1}^{l} score(s[i:])}{l}$

Построение таблицы CT

Вход: множество текстов и коллекция строк

- Каждый текст представляем АСД
- Оцениваем степень сходство каждой строки каждлом АСД

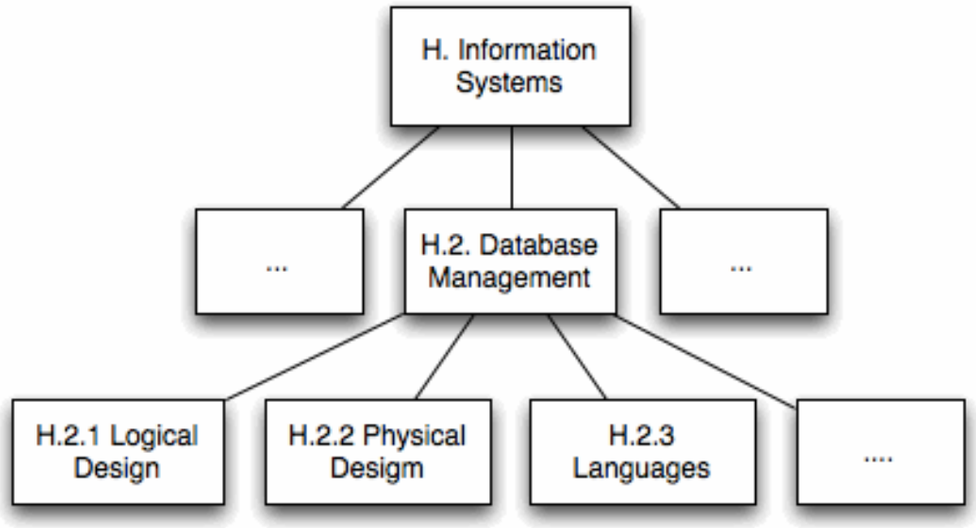
Выход: таблица CT

	Text1	Text2	Text3
String1			
String2			

Приложение 1: аннотирование статьи таксономическими темами – I

Article: Two variable logic on data trees and XML reasoning, Journal of the ACM, 2003

Motivated by reasoning tasks for XML languages, the satisfiability problem of logics on data trees is investigated. The nodes of a data tree have a label from a finite set and a data value from a possibly infinite set. It is shown that satisfiability for two-variable first-order logic is decidable if the tree structure can be accessed only through the child and the next sibling predicates and the access to data values is restricted to equality tests. From this main result, decidability of satisfiability and containment for a data-aware fragment of XPath and of the implication problem for unary key and inclusion constraints is concluded.Motivated by reasoning tasks for XML languages, the satisfiability problem of logics on data trees is investigated. The nodes of a data tree have a label from a finite set and a data value from a possibly infinite set. It is shown that satisfiability for two-variable first-order logic is decidable if the tree structure can be accessed only through the child and the next sibling predicates and the access to data values is restricted to equality tests. From this main result, decidability of satisfiability and containment for a data-aware fragment of XPath and of the implication problem for unary key and inclusion constraints is concluded.



Приложение 1: аннотирование статьи таксономическими темами – II

- **Вход**:
  - таксономия ACM-CCS
  - коллекция аннотаций статей из журнала ACM
- **Построения:** CT таблица таксономическая_тема X аннотация_статьи
- **Найти:** профиль каждой статьи
  - в профиль статьи включаем таксономические темы с высокими оценками

Входные данные

Collection of the ACM Journal abstracts	The ACM Computing Classification System (1998)
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 Mikoaj Bojańczyk, Anca Muscholl, Thomas Schwentick, Luc Segoufin Article No.: 13 doi>10.1145/1516512.1516515 Full text: PDF	… ▪ D. Software ▪ D.0 GENERAL ▪ D.1 PROGRAMMING TECHNIQUES (E) ▪ D.1.0 General ▪ D.1.1 Applicative (Functional) Programming ▪ D.1.2 Automatic Programming (I.2.2) ▪ D.1.3 Concurrent Programming ▪ *Distributed programming* ▪ *Parallel programming* …

Пример “хорошего” ACT-профиля

Article: Two variable logic on data trees and XML reasoning, Journal of the ACM, 2003					
AST found profile			ACM-CCS index terms (manual annotation)		
ID	TE	ACM–CCS topic	ID	#	ACM–CCS topic
H.2.3	0.4541	Languages	H.2.3	0	Languages
I.1.3	0.4489	Languages and Systems	F.4.3	2	Formal Languages
F.4.3	0.3918	Formal Languages	H.2.1	12	Logical Design
D.4.5	0.3049	Reliability	F.4.1	27	Mathematical Logic
I.6.2	0.2578	Simulation Languages	I.7.2	52	Document Preparation

Пример “плохого” ACT-профиля

Article: Lower bounds for processing data with few random accesses to external memory, Journal of the ACM, 2003					
AST found profile			ACM-CCS index terms (manual annotation)		
ID	TE	ACM–CCS topic	ID	#	ACM–CCS topic
H.2.8	0.4330	Database Applications	F.1.3	160	Complexity Measures and Classes
H.2.5	0.2904	Heterogeneous Databases	H.2.4	165	Systems
C.5.1	0.2630	Large and Medium (`Mainframe'') Computers	F.1.1	219	Models of Computation
J.1	0.2115	ADMINISTRATIVE DATA PROCESSING			
I.2.7	0.1870	Natural Language Processing			

Приложение 2: построение таксономии на основе ресурсов Википедии

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

- **Построения:** три CT таблицы таксономическая_тема X название_категории, название_категории X название_статьи, название_родительской_категории X название_статьи

- **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

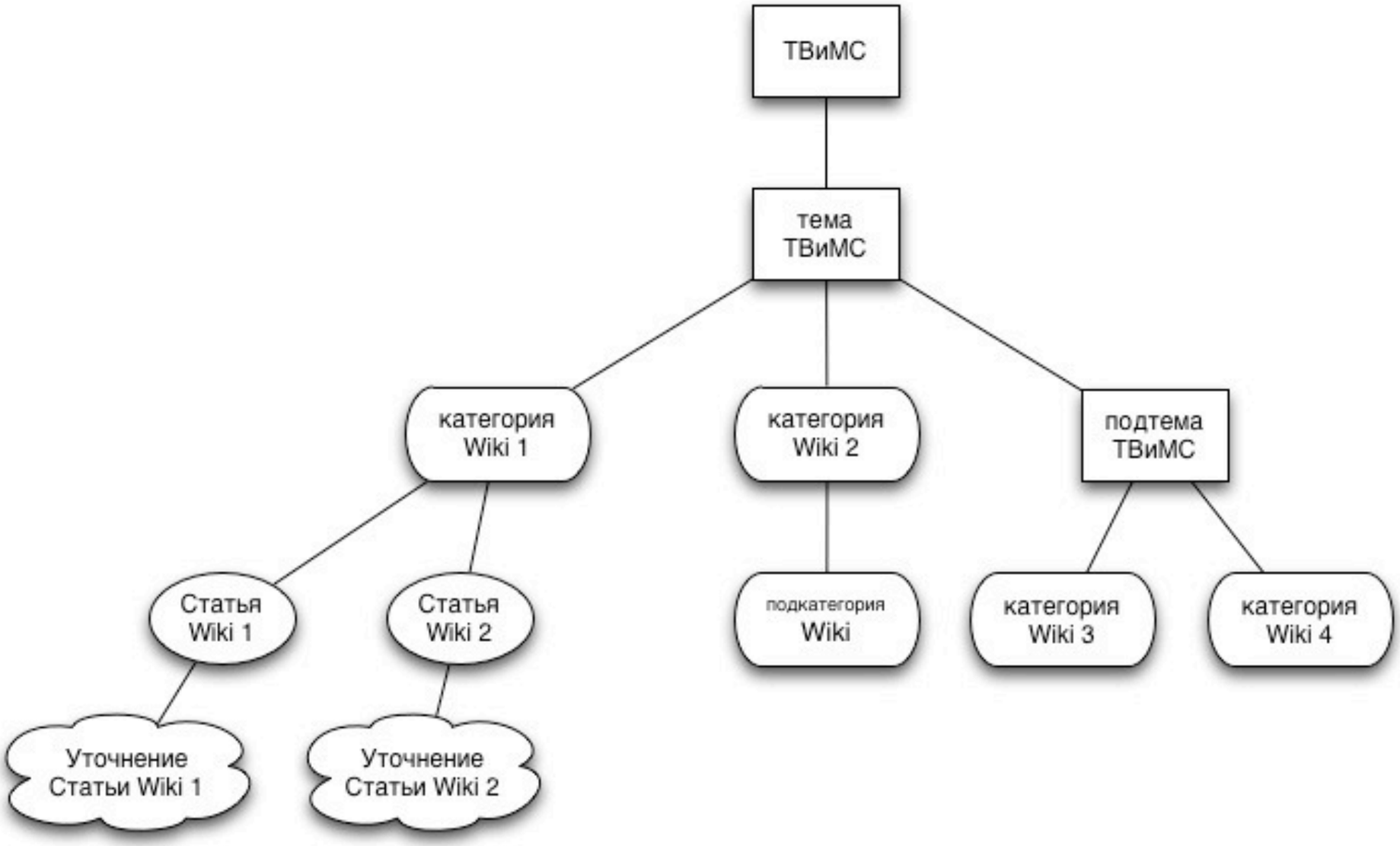
Факторный анализ — многомерный метод, применяемый для изучен количества неизвестных переменных и случайной ошибки.

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

Использование метода АСД в задаче достраивания таксономии

1. Для очистки дерева категорий от иррелевантных статей:

   вычисление степени выраженности названия категории в тексте статьи

2. Для очистки дерева категорий от иррелевантных категорий

   вычисление степени выраженности названия категории в совокупности статей данной категории

3. Для достраивания категорий к промежуточным уровням таксономии

   вычисление степени выраженности таксономических тем в совокупности статей данной категории

Очистка дерева категорий от иррелевантных статей

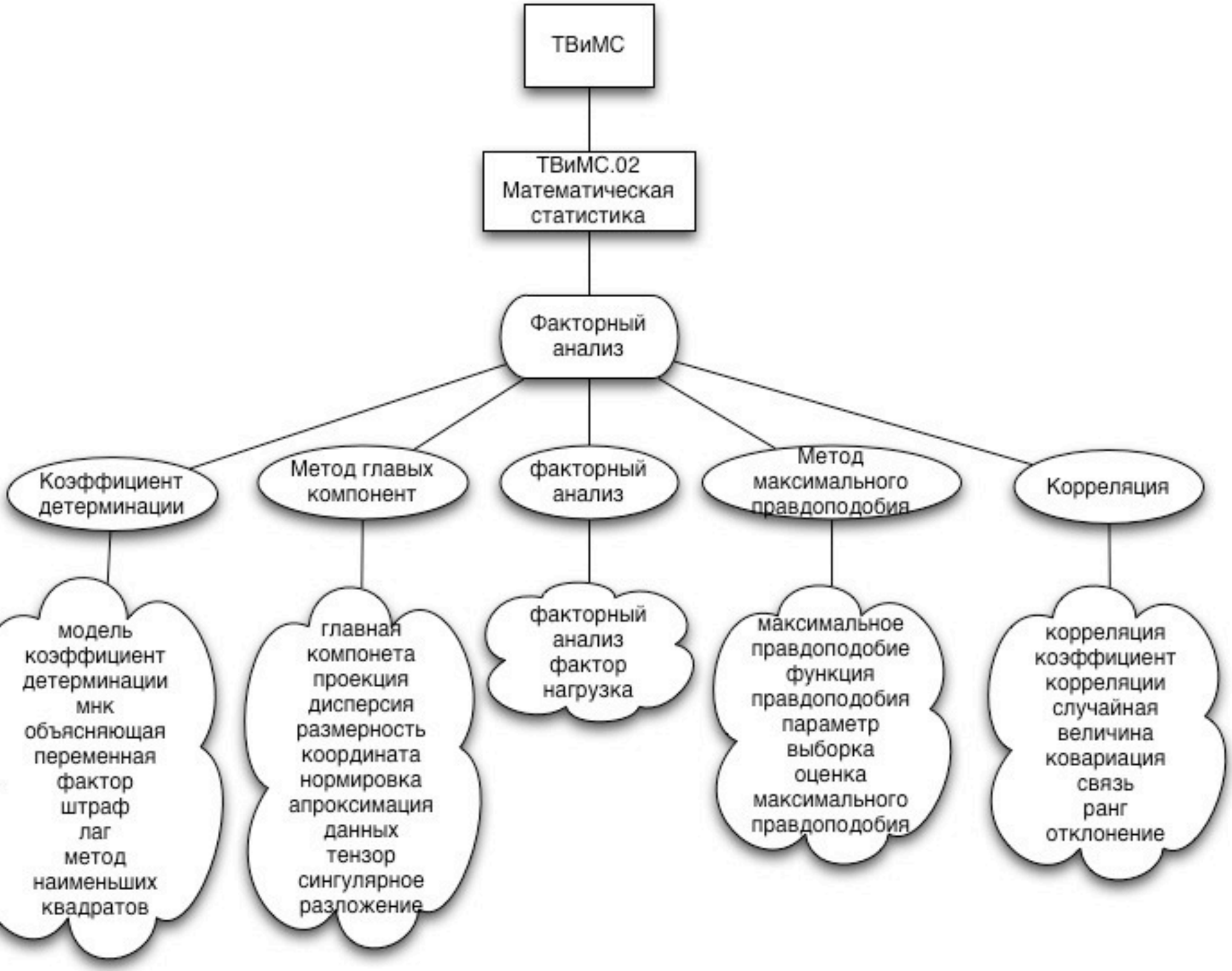
Статьи категории Факторный анализ	
факторный анализ	0.561
метод максимального правдоподобия	0.529
корреляция	0.349
коэффициент детерминации	0.231
метод главных компонент	0.207
линейная регрессия на корреляции	0.157
коррелятор	0.143
RANSAC	0.097
Робастность в статистике	0.067

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

Приложение 3: построение графа связей между ключевыми словосочетаниями – I

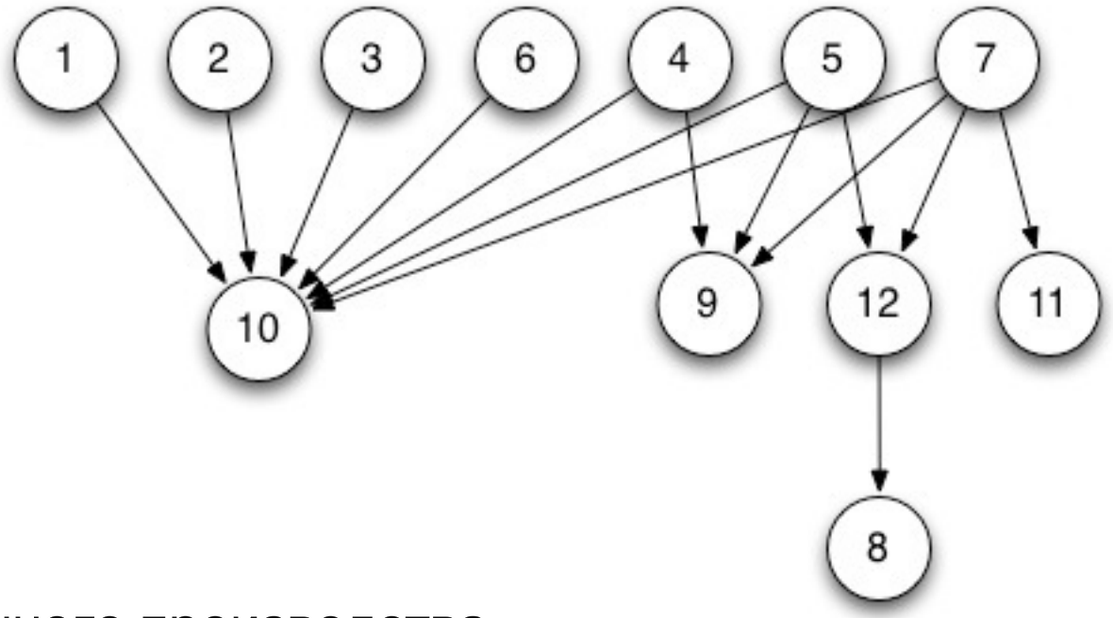
- **Вход**:
  - коллекция веб-публикаций о бизнес-процессе в после-кризисной России
  - множество ключевых словосочетаний, характеризующих типовые события (“публикация финансовой отчетности”, “смена генерального директора”)
- **Построения:** CT таблица ключевое_словосочетание X веб_публикация
- **Найти:** ключевое_словосочетание -> множество публикаций
  - веб-публикация относится к ключевому словосочетанию, если мера релевантности ключевого сочетания данной веб-публикации выше заданного порога

Приложение 3: построение графа связей между ключевыми словосочетаниями – II

Импликации на ключевых словосочетаниях

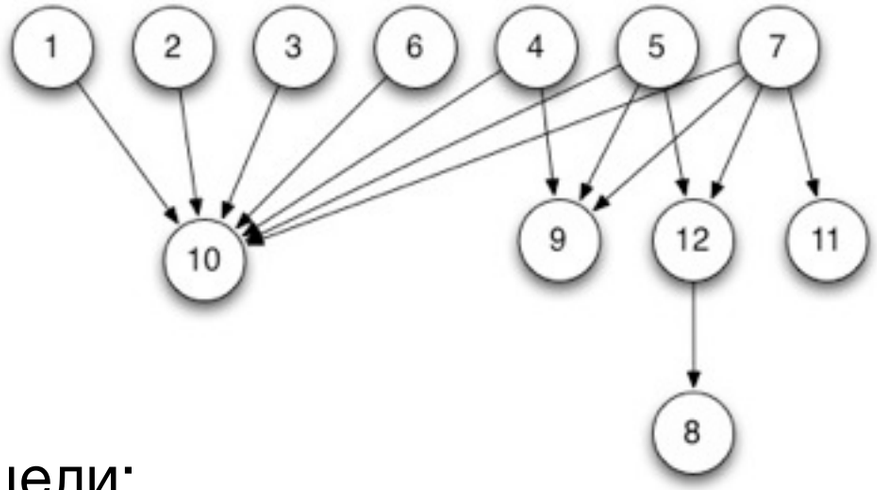
- F(A) – подмножество веб-публикаций, относящихся к ключевому словосочетанию A
- F(B) – подмножество веб-публикаций, относящихся к ключевому словосочетанию B
- из A следует B, если доля F(B) в F(A) > 60%
- **Пример**: Ввод автоматизированного производства ⇒ Повышение эффективности управления завтратами

Граф связей между ключевыми словосочетаниями по коллекции текстовых документов



1. Ввод автоматизированного производства.
2. Выпуск пресс-релизов (с положительными или отрицательными новостями).
3. Изменение размера пакета акций, принадлежащего институциональному инвестору.
4. Изменение уровня концентрации собственности.
5. Повышение квалификации персонала.
6. Проведение вертикального слияния.
7. Проведение операций купли-продажи бренда.
8. Выход на международный рынок.
9. Изменение организационно-правовой формы.
10. Повышение эффективности управления затратами.
11. Публикация финансовой отчетности.
12. Смена финансового директора.

Graph of the interrelation between the key phrases over the text collection



Показанные в графе цели:

- уменьшение издержек (10),
- изменение организационно- правовой формы (9),
- повышение прозрачности (11),
- выход на мировые рынки (8).

Событие (12) оказывается шагом, ведущим к выходу на мировые рынки.

Основные факторы бизнес-процесса:

- купля-продажа брендов (развитие сетевых структур),
- автоматизация производства,
- повышение квалификации персонала,
- передача государственных активов в частные руки.

Conclusion and discussion

- Interpretation by producing profiles and lifting them in the taxonomy

- Issues

A. AST scoring – slow and noised (Future work: some linguistics preprocessing)

B. The taxonomies are not quite relevant

C. Penalty weights? (Future work: change the parsinomy criterion for that of the maximum likelihood)

D. Assessment of the results

Будущая работа

Дальшейшее развитие метода

- Автоматическое извлечение ключевых словосочетаний произвольной длины
- Автоматическое порождение и использование синонимов и околосинонимов

Данная работа выполнялась при частичной финансовой поддержке Научного Фонда НИУ-ВШЭ через коллективный исследовательский проект «Учитель-Ученики» 11-04-0019.