Алгоритмы извлечения из неструктурированных текстовых источников метаинформации о научно-технических конференциях

А.В. Бахтин

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М.В. ЛОМОНОСОВА
Механико-математический факультет

Москва, 2015

Общий план

Мотивация

Информация о деятельности мирового научного сообщества является важной при решении ряда аналитических задач, таких как:

- определение скорости развития и уровня интереса к тому или иному направлению научных исследований;
- выделение и проведение анализа актуальных и зарождающихся направлений научных исследований.

Показатели такого рода могут быть использованы:

- руководителями научных организации или групп для более объективного анализа результатов работы;
- в качестве отправной точки для различных исследований, например, по социологии или философии науки.

Мотивация

Анализ активности научного сообщества предполагает наличие некоторого источника информации о работе его членов письменных источников.
Возможно анализировать публикации:

- книги;
- статьи в журналах и трудах конференций.

Объектом анализа может быть как текст, так и данные о цитировании.

Недостатки традиционного подхода:

- недостаточная оперативность;
- консервативность в выборе тематики.

Сообщения о конференциях лишены недостатков публикаций и позволяют оценивать актуальность научных направлений в ближайшем будущем.

Мотивация

Информация о прошедших и планируемых конференциях имеет и самостоятельный интерес.

- Наличие базы знаний о датах, тематике и месте проведения конференций позволит создать систему с удобными механизмами поиска и навигации по предстоящим событиям, которая будет учитывать предпочтения пользователя о месте проведения и область его интересов.
- Данные о программном комитете конференции в совокупности с внешними данными об индексе цитируемости его участников могут служить основой для автоматического построения рейтинга конференции.
  - поиск наиболее авторитетных конференций в смежных областях для учёных;
  - оценка научной работы сотрудников специализированными аналитическими системами.

Описание задачи

Задача состоит в создании средств извлечения метаинформации о конференциях, включающей в себе данные о дате и месте проведения конференции, ключевых тематиках конференции и программном комитете.

Обзор работ

Классические работы по извлечению информации предполагают наличие вербального контекста, отсутствующего в сообщения о конференциях.

Множество работ по извлечению данных о конференциях можно разделить на несколько категорий по типу источников данных:

- работы с использованием ручного ввода с заполнением нескольких полей (ACM Calendar, WikiCFP, EventSeer);
- работы, использующие данные с домашних страниц конференций (Xin CIKM'08);
- работы, использующие автоматические способы анализа текстовых источников для извлечения ограниченного набора полей (Ireson ICML'05, Schneider 2007, Amaro CISTI'10, Li CIKM'13).

Пример объявления о конференции

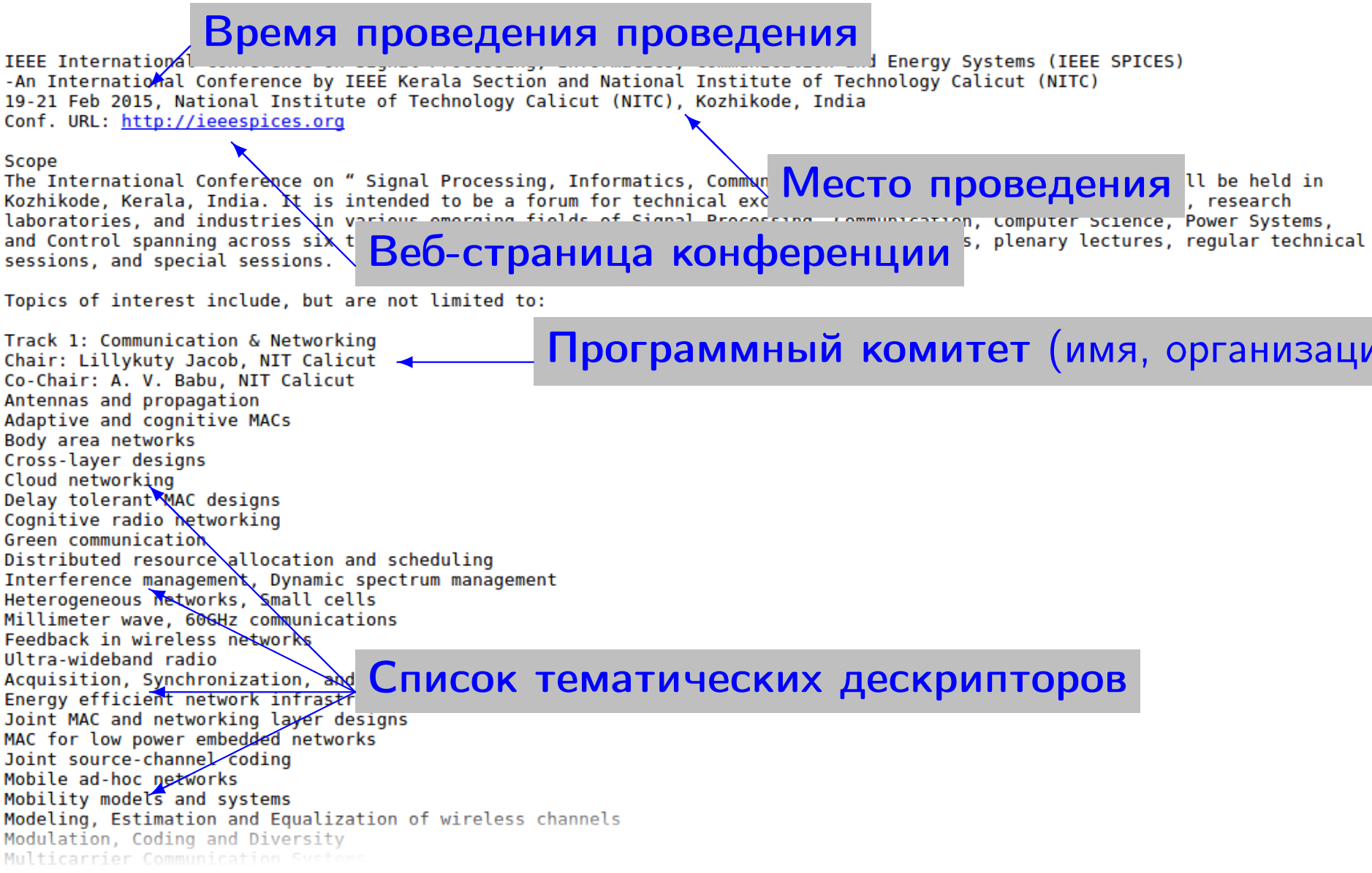
IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (IEEE SPICES)
-An International Conference by IEEE Kerala Section and National Institute of Technology Calicut (NITC)
19-21 Feb 2015, National Institute of Technology Calicut (NITC), Kozhikode, India
Conf. URL: http://ieeespices.org

Scope
The International Conference on " Signal Processing, Informatics, Communication and Energy Systems (SPICES)" will be held in Kozhikode, Kerala, India. It is intended to be a forum for technical exchange amongst researchers from academia, research laboratories, and industries in various emerging fields of Signal Processing, Communication, Computer Science, Power Systems, and Control spanning across six tracks. The technical program will include keynote lectures, plenary lectures, regular technical sessions, and special sessions.

Topics of interest include, but are not limited to:

Track 1: Communication & Networking
Chair: Lillykuty Jacob, NIT Calicut
Co-Chair: A. V. Babu, NIT Calicut
Antennas and propagation
Adaptive and cognitive MACs
Body area networks
Cross-layer designs
Cloud networking
Delay tolerant MAC designs
Cognitive radio networking
Green communication
Distributed resource allocation and scheduling
Interference management, Dynamic spectrum management
Heterogeneous networks, Small cells
Millimeter wave, 60GHz communications
Feedback in wireless networks
Ultra-wideband radio
Acquisition, Synchronization, and Tracking
Energy efficient network infrastructure
Joint MAC and networking layer designs
MAC for low power embedded networks
Joint source-channel coding
Mobile ad-hoc networks
Mobility models and systems
Modeling, Estimation and Equalization of wireless channels
Modulation, Coding and Diversity
Multicarrier Communication Systems

Пример объявления о конференции



IEEE International [illegible] Energy Systems (IEEE SPICES)
-An International Conference by IEEE Kerala Section and National Institute of Technology Calicut (NITC)
19-21 Feb 2015, National Institute of Technology Calicut (NITC), Kozhikode, India
Conf. URL: http://ieeespices.org

Scope
The International Conference on " Signal Processing, Informatics, Commun[illegible]ll be held in
Kozhikode, Kerala, India. It is intended to be a forum for technical exc[illegible], research
laboratories, and industries in various emerging fields of Signal Processing, Communication, Computer Science, Power Systems,
and Control spanning across six t[illegible]s, plenary lectures, regular technical
sessions, and special sessions.

Topics of interest include, but are not limited to:

Track 1: Communication & Networking
Chair: Lillykuty Jacob, NIT Calicut
Co-Chair: A. V. Babu, NIT Calicut
Antennas and propagation
Adaptive and cognitive MACs
Body area networks
Cross-layer designs
Cloud networking
Delay tolerant MAC designs
Cognitive radio networking
Green communication
Distributed resource allocation and scheduling
Interference management, Dynamic spectrum management
Heterogeneous Networks, Small cells
Millimeter wave, 60GHz communications
Feedback in wireless networks
Ultra-wideband radio
Acquisition, Synchronization, and
Energy efficient network infrastr
Joint MAC and networking layer designs
MAC for low power embedded networks
Joint source-channel coding
Mobile ad-hoc networks
Mobility models and systems
Modeling, Estimation and Equalization of wireless channels
Modulation, Coding and Diversity
Multicarrier Communication Systems

Пример объявления о конференции

```
**************  SIGMOD 2016 Call for Contributions ************************
NEW:   Early first deadline this year: July 9, 2015
***************************************************************************
        ACM SIGMOD International Conference on Management of Data
                Hyatt Regency, San Francisco, California, USA
                                        June 26 - July 1, 2016

                                        http://www.sigmod2016.org

                                    https://twitter.com/acmsigmod2016
                                    https://www.facebook.com/sigmod2016

***************************************************************************

The annual ACM SIGMOD conference is a leading international forum for database researchers,
practitioners, developers, and users to explore cutting-edge ideas and results, and to
exchange techniques, tools, and experiences. It is co-located with PODS focusing on the
theoretical aspects of databases.

ACM SIGMOD 2016 solicits submissions for the following programs and events:
* Research papers
* Technical demonstrations
* Industrial papers
* Tutorials
* Panels
* Workshops

We invite the submission of original contributions relating to all aspects of data
management defined broadly, and particularly encourage submissions on topics of emerging
interest in the research and development communities.
```

Пример объявления о конференции

********** SIGMOD 2016 Call for Contributions **********
NEW: Early first deadline this y
****

ACM SIGMOD International Conference on Management of Data
Hyatt Regency, San Francisco, California, USA
June 26 - July 1, 2016

http://www.sigmod2016.org

https://twitter.com/
https://www.facebook

Место проведения

Время проведения

Веб-страница конференции

****

The annual ACM SIGMOD conference is a leading international forum for database researchers, practitioners, developers, and users to explore cutting-edge ideas and results, and to exchange techniques, tools, and experiences. It is co-located with PODS focusing on the theoretical aspects of databases.

ACM SIGMOD 2016 solicits submissions for the following programs and events:
* Research papers
* Technical demonstrations
* Industrial papers
* Tutorials
* Panels
* Workshops

We invite the submission of original contributions relating to all aspects of data management defined broadly, and particularly encourage submissions on topics of emerging interest in the research and development communities.

Постановка задачи

Определение

Сообщение о конференции (информационное письмо, CFP) — текст, содержащий основную информацию о планируемой конференции.

Определение

Фрейм — структура представления данных, состоящая из названия и фиксированного количества полей.

Задача

По текстовому сообщению о конференции заполнить фрейм о конференции.

Решение данной задачи подразумевает создание алгоритмов извлечения, демонстрирующих высокую точность извлечения на контрольной выборке.

Общий план

Архитектура системы заполнения фрейма

Основа системы — модель документа, которая хранит текст в виде списка токенов и извлечённую из текста разметку (элементы списков, страны, даты и.т.д).

Работа системы состоит в последовательном выполнении правил, каждое из которых принимает модель документа, и возвращает модифицированную модель документа.

Модель документа

Определение

Модель документа — есть пара (T, M), состоящая их списка токенов и разметки.

Определение

Список токенов — отображение $T : \{1, 2, \ldots, n\} \to \mathfrak{T}$, где n — длина текста в токенах, а $\mathfrak{T}$ — множество всех возможных токенов. Каждый токен суть тройка $(\Sigma^+, type, offset)$, где Σ — используемый алфавит текста, $type \in \{word, punct, space, nl\}$ — тип токена, а $offset$ — позиция токена в тексте.

Определение

Разметка суть конечное множество плиток $M = \{Tile_1, Tile_2, \ldots, Tile_k\}$

Определение

Плитка — четвёрка $(position_{start}, position_{end}, type, attrs)$, где:

- $position_{start}$ и $position_{end}$ — позиции соответственно первого и последнего накрываемых токенов соответственно;
- $type \in \mathfrak{E}$ — тип плитки из конечно множества типов $\mathfrak{E}$;
- $attrs$ — словарь атрибутов.

Выделение дат и страны проведения

Выделение даты начала и даты окончания конференции:

- правило выделения временных объектов (регулярные выражения);
- правило разрешения конфликтов вложенности;
- правило выбора даты начала и даты окончания конференции (оценка позиции и контекста).

Выделение страны проведения конференции.

- правило поиск названий стран (поиск по газитиру);
- правило выбора страны проведения (оценка позиции).

Выделение домашней страницы

Выделение домашней страницы:

- правило выделения электронных адресов (URL-адреса и адреса электронной почты);
- правило выбора адреса домашней страницы.

В качестве домашней выбирается адрес, ближайший к упоминанию страны в тексте.

Выделение тематики конференции

Список тематик конференции определяет перспективные научные направления, обсуждению и работе над которыми посвящена конференций.

В тексте сообщения о конференции тематика конференции представлена в виде списка **тематических дескрипторов**.

Пример тематических дескрипторов в CFP

WER'13 — Workshop on Requirements Engineering

WER'13 is the sixteen edition of the Workshop on Requirements Engineering. The worksh
of researchers from other parts of the world as well. WER\u201913 will be carried out

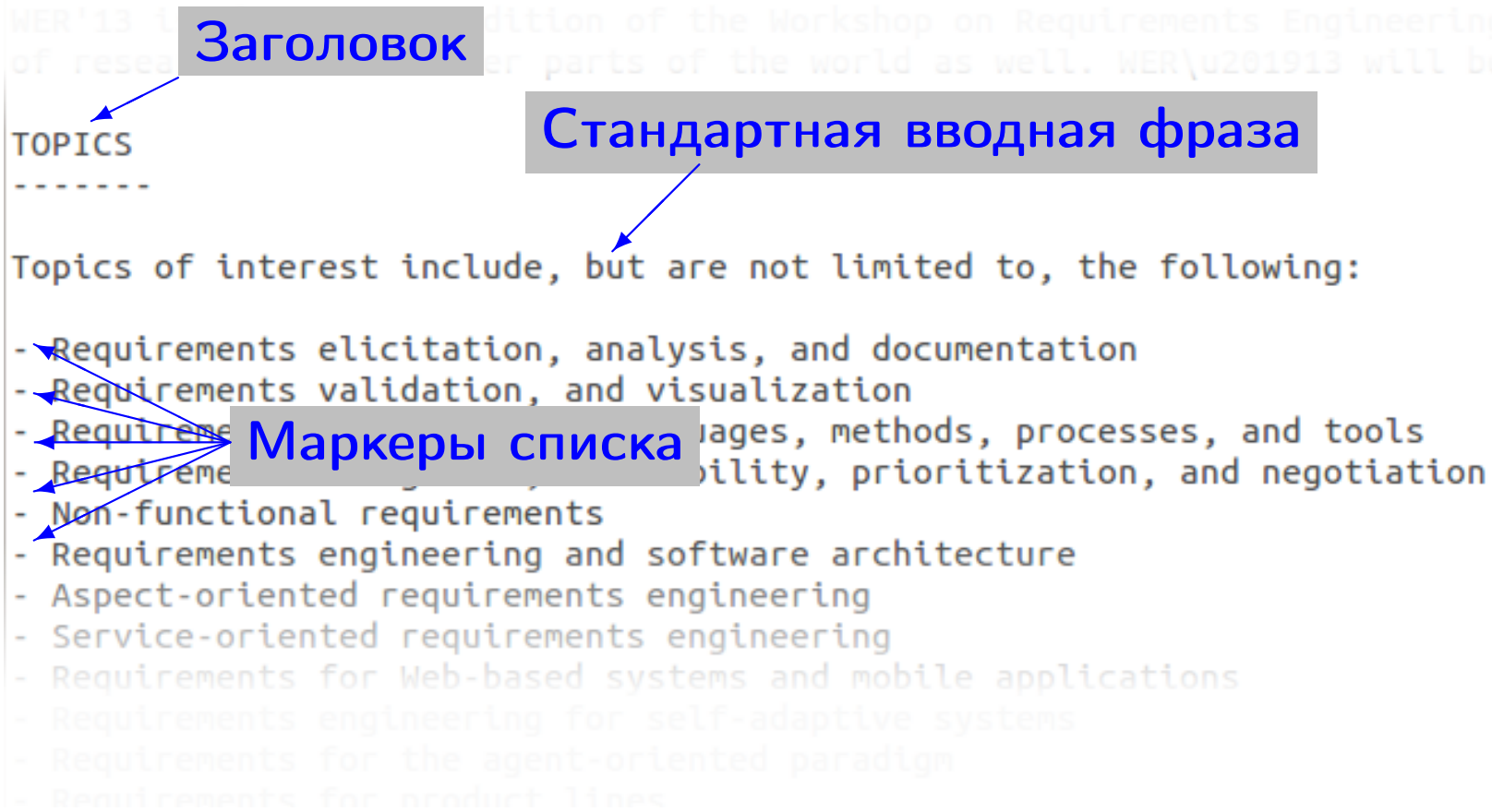
```
TOPICS
-------

Topics of interest include, but are not limited to, the following:

- Requirements elicitation, analysis, and documentation
- Requirements validation, and visualization
- Requirements specification languages, methods, processes, and tools
- Requirements management, traceability, prioritization, and negotiation
- Non-functional requirements
- Requirements engineering and software architecture
- Aspect-oriented requirements engineering
- Service-oriented requirements engineering
- Requirements for Web-based systems and mobile applications
- Requirements engineering for self-adaptive systems
- Requirements for the agent-oriented paradigm
- Requirements for product lines
```

Пример тематических дескрипторов в CFP

WER'13 — Workshop on Requirements Engineering



Наивный алгоритм выделения тематических дескрипторов

- **Выделение списков**
  поиск строк, первое слово которых — один и тот же символ пунктуации
- **Выбор списков, содержащих тематические дескрипторы**
  поиск заголовка в предшествующей строке

Примеры, не обрабатываемые наивным алгоритмом

ALAIPO — 2Latin Association of Human-Computer Interaction

Нетривиальная вводная фраза

areas are solicited on, but not limited to:

Маркеры, содержащие более одного символа

:: Advances in Human-Computer Interface
:: Aesthetic and Creative Design
:: Archit
:: Auditory Contents for Interactive Systems
:: Biometrics Techniques and Privacy
:: Cloud Computing
:: Cognitive Modeling
:: Collaborative Learning
:: Communicability in Hypertext, Multimedia and Hypermedia Syste
:: Community and Security Management of Emerging Networks and Se
:: Computer Animation: 2D, 3D, and N-D Animation Systems

Примеры, не обрабатываемые наивным алгоритмом

```
ICEL2013 welcomes submissions on any topic in the field of e-learning.

The conference welcomes papers on the following (but not limited to)
research topics:

+ Architecture of Educational Information Systems Infrastructure
+ Education for Computer-mediated
+ Electronic L
+ e-Learning M                                          ods and Approaches
+ e-Learning Tactics, Pedagogical Strategies, Curriculum Development
Issues
+ e-Moderating, e-Tutoring, and e-Facilitating
+ e-Skills and Information Literacy for Learning
+ Instructional Design
+ ICT Skills Education and Online Assessment
+ Learning Management Systems (LMS)
```

Принудительные разрывы строк

Примеры, не обрабатываемые наивным алгоритмом

CAISE 2011 — 23rd International Conference on Advanced Information Systems

year, the conference will be particularly glad to receive exploratory papers that share visions, novel approaches, or insight into new information systems engineering paradigms. The CAiSE topics of interests include, but are not restricted to:

Methodologies and Approaches for IS Engineering:
- Innovation and creativity in IS engineering
- Enterprise architecture and enterprise modelling

<некоторое количество строк>

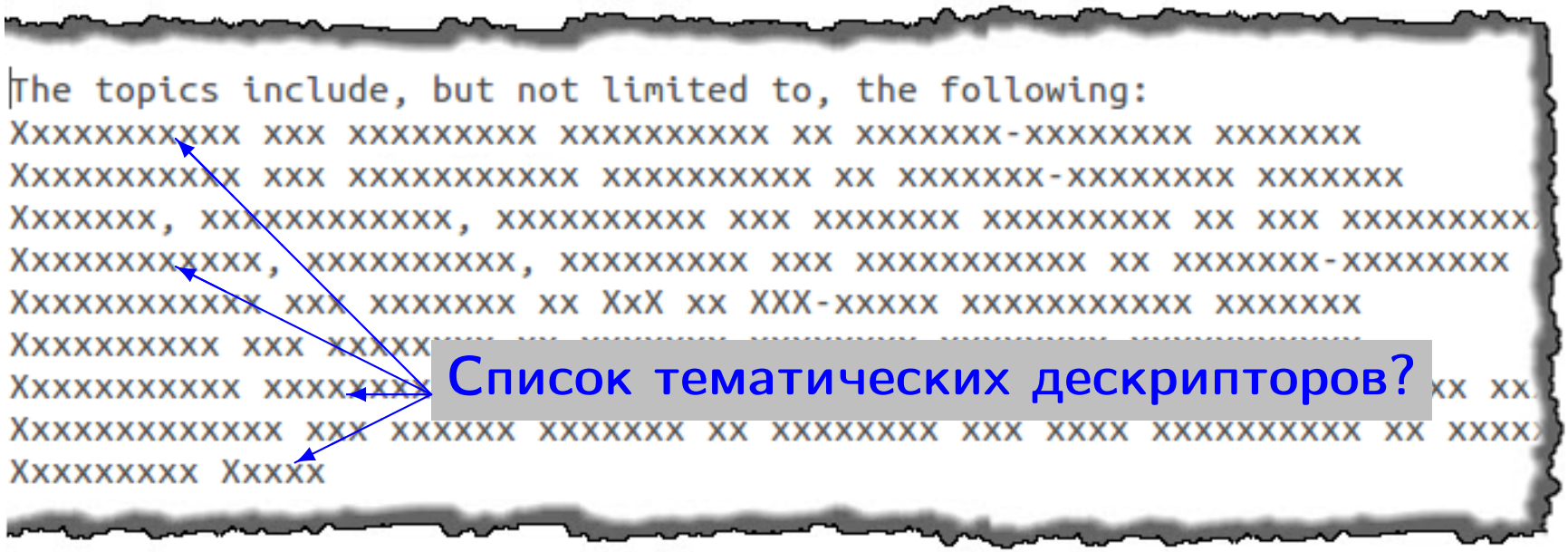
- Service science

Innovative platforms, architectures and technologies for IS:
- Service-oriented architecture
- Model-driven architecture
- Component based development
- Agent architecture
- Distributed, mobile, and open architecture

Отсутствие вводной строки

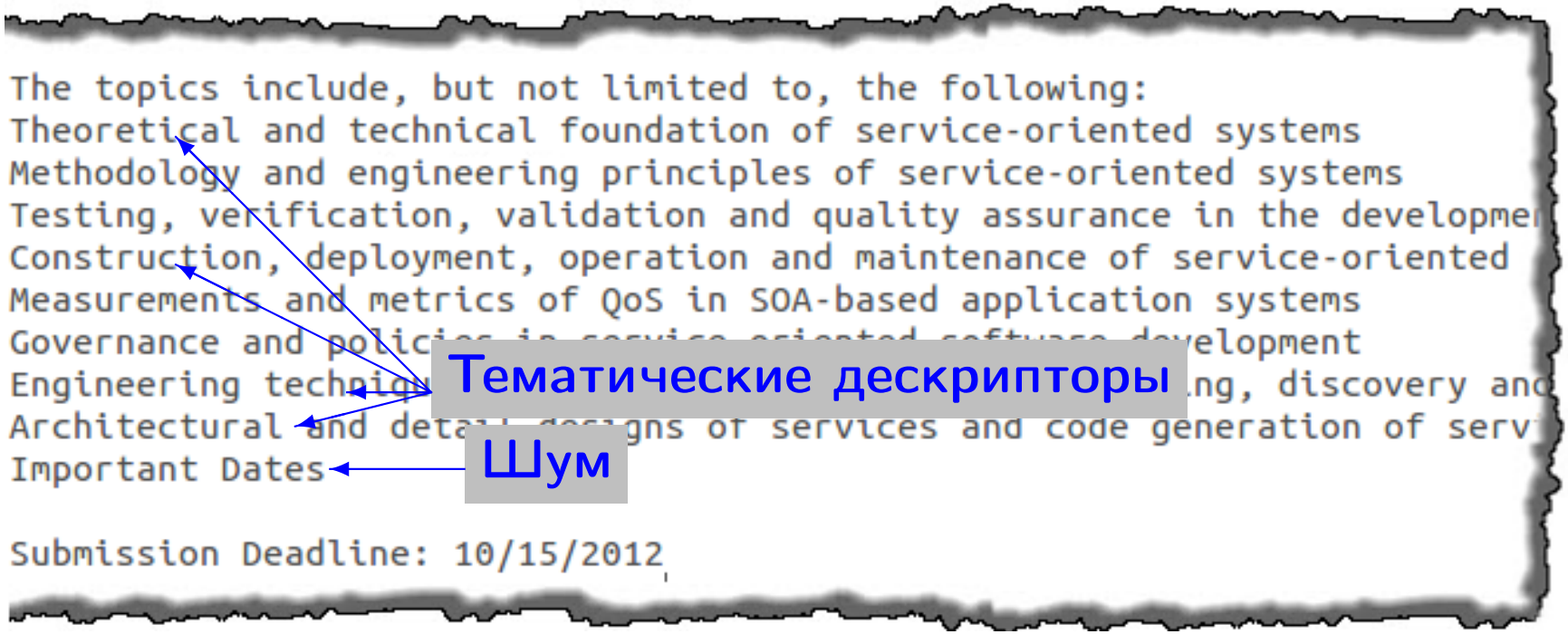
Более сложные примеры

SOSE 2013 — 7th International Symposium on Service Oriented System Engineering



Более сложные примеры

SOSE 2013 — 7th International Symposium on Service Oriented System Engineering



Примеры, не обрабатываемые наивным алгоритмом

Плоские списки: все элементы списка записаны в одну строку.

- Xxx xxxxxxxx xxxxxxxx xxxx xxxxx, xxxxxxx xxx xxxxxxxxxx xxxxxxxxx xx xxxxxxxx xxxxxxxx xxxxxxxxxx xx xxxxxxxx, xxxxxxxx xxx xxxxxxxx.
- Xxx Xxxxxxxxxxxxx Xxxxxxx Xxxxxxxxx xxxx xxxxxx xxx xxxxx xx xxxxxx-xxxxx xxxxxx: xxxxxxxx xxxxx (10 xxxxx, 2 xxxxxxx); xxxxxxx xxxxx (6 xxxxx, 2 xxxxxxx).
- Xxxx Xxxxxxxxx, Xxxxxxxxx Xxxxxxxxx, Xxxxxxx Xxxxxxxx, Xxxxxxxxxx Xxxxxxxxxx, Xxxxxxxxxx Xxxxx, Xxxxxxxxxx Xxxxxxxxxx.

Примеры, не обрабатываемые наивным алгоритмом

Плоские списки: все элементы списка записаны в одну строку.

- `The audience includes both users, vendors and scientists operating in applying advanced techniques to industry, business and services.`
- `The International Program Committee will accept two types of camera-ready papers: extended paper (10 pages, 2 columns); regular paper (6 pages, 2 columns).`
- `Bulk Terminals, Container Terminals, Harbour Services, Industrial Facilities, Navigation Lines, Multimodal Transports.`

Более сложный алгоритм

1. Исправление текста
2. Выделение списков
3. Идентификация списков, содержащих тематические дескрипторы

Исправление текста

Необходимо для каждого символа переноса строки \n определить, несёт ли он семантическую нагрузку.

```
studies on e-government from developed and developing countries will\n
\n
be encouraged.\n
\n
Contributed papers may deal with, but are not limited to:\n
\n
*    Theories and conceptual models informing citizens adoption of e-government\n
\n
*    Theories and models of citizens' satisfaction with e-government\n
\n
implementations\n
\n
*    Citizens participation in e-democracy and e-governance\n
\n
*    Cross country comparison of citizens perspectives on e-government\n
```

Исправление текста

Необходимо для каждого символа переноса строки \n определить, несёт ли он семантическую нагрузку.

```
studies
\n
be enc
\n
Contri
\n
*    Th                                                    vernment\n
\n
*    Th
\n
impleme
\n
*    Ci
\n
*    Cr                                                    n
```

Что влияет на принятие решения?

- статистики длин строк (максимальная, минимальная, концентрация, разница средней и текущей);
- положение строки в файле;
- статистики количества переносов (максимальное, среднее, отношение к текущему);
- статистические данные о вероятности окончания предложения словом;
- наличие признаков структуры (пробельный префикс, общий префикс с предыдущей строкой).

Используя эти факторы был построен классификатор.

Выделение списков. Типы списков

- **Списки с маркерами**
  кластеризация префиксов слов;
- **Плоские списки**
  множество правил с порогами *качественности* для каждого типа разделителя;
- **Списочные области**
  интегральное правило для списков без маркеров.

Выделение списков. Списочные области

Задача: найти строки, относящиеся к какому-либо списку

engineering, computer programming, social sciences, advocacy, doctors
in psychiatry, etc. from production, circulation, safety, security,
etc. of the digital information to control and the prevention of the
cyber attacks.

Scopes of Interest (not limited to):

Access Control and Privacy Protection
Access to Information and Knowledge: Rights and Obligations
Advanced Multimodal Interfaces for Security

<некоторое количество строк>

Web 2.0 and Web 3.0: Safety, Security and Privacy
Wireless Data Communications: Security and Privacy
Submissions

Two-stage submission: First, interested researchers and practitioners
are invited to submit a chapter proposal clearly stating your focused
domain problems and contributions related to one of the above topics

Выделение списков. Списочные области

Задача хорошо описывается в терминах последовательной маркировки

Выделение списков. Списочные области

Формальное описание графического описания модели случайных условных полей.

- Прямоугольники соответствуют случайным величинам одного из двух типов:
  - $m_i \in \{0, 1\}$ — относится ли i-я строка к списочной области или нет;
  - $\overline{X}_i$ — признаковое описание i-й строки.
- Величины прямоугольников, не соединённые ребром, являются условно-независимыми.

Таким образом, совместная вероятность имеет следующий вид:

$$\mathrm{P}(m_1, \overline{X}_1, m_2, \overline{X}_2, \ldots, m_n, \overline{X}_n) = \frac{1}{Z} \prod_i f(m_i, m_{i+1}) g(m_i, \overline{X}_i),$$

где $f(\cdot)$ и $g(\cdot)$ — некоторые неотрицательные функции, а Z — нормирующая константа.

Выделение списков. Списочные области: факторы

1. Строка содержит маркер списка
2. Строка содержит более двух слов
3. Строка в верхнем регистре
4. Строка содержит имя
5. Строка имеет пробельный префикс
6. Строка заканчивается двоеточием
7. Первый токен совпадает с первым токеном i-й строки
8. Последний токен совпадает с последним токеном i-й строки
9. Строка содержит вводную строку для топиков
10. Строка содержит тематический дескриптор

Откуда брать список тематических дескрипторов?

- Внешние данные (например, заголовки статей Википедии)
  - малая полнота;
  - консервативность.

Откуда брать список тематических дескрипторов?

- Внешние данные (например, заголовки статей Википедии)
  - малая полнота;
  - консервативность.
- Ручная разметка дескрипторов в имеющихся данных
  - малая эффективность.

Откуда брать список тематических дескрипторов?

- Внешние данные (например, заголовки статей Википедии)
  - малая полнота;
  - консервативность.
- Ручная разметка дескрипторов в имеющихся данных
  - малая эффективность.
- Автоматическая разметка дескрипторов в имеющихся данных
  - двухпроходная архитектура: выбираем дескрипторы, которые получены разметкой *простых* случаев;
  - на тестовой выборке увеличение полноты на 10%.

Обнаружение списков тематик

Необходимо пометить списки, содержащие ключевые тематики.

- Маркеры предсписочного текста
- Маркеры заголовка секции, содержащей список
- Метка предыдущего списка
- Антимаркеры предсписочного текста

Список членов программного комитета

Существует стандартный порядок описаний программного комитета: <имя>, <университет>, <страна>.

```
Co-Chairs

Mo-Che Chan, National Central University, Taiwan
Simon Fong, University of Macau, Macau
```

Однако в зависимости от сообщения формат может быть произвольно изменён:

```
Program Committee:
Sihem Amer-Yahia                    CNRS LIG, France
Gildas Avoine                            UCL, Belgium
Boualem Benatallah                  UNSW, Australia
Gregor V. Bochmann                Univ of Ottawa, Canada
```

Для решения задачи в общем случае необходимо (с некоторой точностью) знать области, которые похожи на имена и организации.

Использование внешних модулей

Модель документа позволяет использовать данные от внешних модулей, которые представлены как размеченные токены.

Ввод	**Вывод**
Mo-Che Chan, National Central University, Taiwan	Mo-Che/O Chan/O ,/O National/ORG Central/ORG University/ORG ,/O Taiwan/LOC
Jeannette E. Riley (Women's Studies)	Jeannette/PERSON E./PERSON Riley/PERSON -LRB-/O Women/O 's/O Studies/O -RRB-/O
–Terry Tempest Williams	--/O Terry/PERSON Tempest/PERSON Williams/PERSON

- Искажение в списке токенов $\Rightarrow$ требуется выравнивание
- Низкая точность $\Rightarrow$ требуются дополнительные данные (газитир)

Задача нечёткого сопоставления пары токенизаций

Пусть имеются токенизации T_1 и T_2 (набор слов $\{w_1^1, w_2^1, \ldots, w_{|T_1|}^1\}$ и $\{w_1^2, w_2^2, \ldots, w_{|T_2|}^2\}$).

Определение (Выравнивание токенизаций)

Выравнивание токенизаций означает построение таких отображений $A_1 : \mathbb{N} \to \mathbb{N} \cup \{0\}$ *и* $A_2 : \mathbb{N} \to \mathbb{N} \cup \{0\}$, *что*

$$A_1(\{1, 2, \ldots, |T_1|\}) \triangle A_2(\{1, 2, \ldots, |T_2|\}) \subset \{0\},$$

и выполнено следующее условие

$$\forall i_1 < i_2 \; \forall j \in \{1, 2\} : (A_j(i_1) \neq 0 \; \wedge \; A_j(i_2) \neq 0) \Rightarrow (A_j(i_1) \leqslant A_j(i_2)).$$

Задача нечёткого сопоставления состоит в построении такого выравнивания (A_1, A_2), которое минимизирует некоторый функционал потерь $\mathcal{L}(T_1, T_2, A_1, A_2)$.

Полный перебор имеет экспоненциальную сложность .

Задача нечёткого сопоставления пары токенизаций

Функция потерь была выбрана на основании того замечания, что внешний модуль не изменяет и не удаляет символы латинских букв.

Она складывается из трёх функций:

- функция штрафа за нарушение измельчения (сопоставление одного слова T_2 нескольких словам T_1);
- функция штрафа за нарушение инъективности сопоставления латинских слов;
- функция штрафа за искажение латинских букв.

Для формализации введём следующие функции и обозначения:

- $|w|$ — длина строки w;
- $is\_space(w)$ — принимает 1, если строка w является обобщённым пробелом, и 0 — иначе;
- $\alpha(\bullet)$ по строке или последовательности строк возвращает все латинские символы, содержащиеся в них;
- $has\_latin(w) := I(|\alpha(w)| > 0)$ — принимает 1, если строка w содержит латинские буквы, и 0 — иначе ($I(\bullet)$ — индикаторная функция);
- $useless(w) := 1 - has\_latin(w)$;
- $full\_latin(w) := I(|\alpha(w)| = |w|)$ — принимает 1, если строка w содержит только латинские буквы, и 0 — иначе;
- $\rho_{ham}(w_1, w_2)$ — расстояние Хемминга между строками w_1 и w_2, т.е. минимальное количество замен символов, необходимое для преобразования строки w_1 к строке w_2.

- функция штрафа за нарушение измельчения:

$$l_1(T_1, T_2, A_1, A_2) = \sum_{\substack{i,j \\ i<j \\ A_1(i)\neq 0}} I(A_1(i) = A_1(j));$$

- функция штрафа за нарушение инъективности сопоставления латинских слов:

$$l_2(T_1, T_2, A_1, A_2) = \sum_{\substack{i \\ full\_latin(w_i^1)>0}} I(A_1(i) = 0 \vee |A_2^{-1}(A_1(i))| \neq 1);$$

- функция штрафа за искажение латинских букв:

$$l_3(T_1, T_2, A_1, A_2) = \sum_{\substack{i,j \\ full\_latin(w_i^1)>0 \\ A_1(i)=A_2(j)}} \rho_{ham}(\alpha(w_i^1), \alpha(w_j^2));$$

Функция потерь:

$$\mathcal{L}_{strict}(T_1, T_2, A_1, A_2) = \sum_{i=1}^{3} \varepsilon_i l_i(T_1, T_2, A_1, A_2),$$

$$\mathcal{L}(T_1, T_2, A_1, A_2) = \mathcal{L}_{strict}(T_1, T_2, A_1, A_2) + \varepsilon_4 l_4(T_1, T_2, A_1, A_2),$$

где ε_1, ε_2, ε_3 и ε_4 — некоторые положительные коэффициенты.

Функция $\mathcal{L}_{strict}(T_1, T_2, A_1, A_2)$ является неотрицательной и принимает нулевое значение в том и только том случае, когда выравнивание удовлетворяет описанным выше неформальным требованиям, но не штрафует за любое другое искажение.

По этой причине среди всех всем сопоставлений, на которых достигается минимум функции $\mathcal{L}_{strict}(T_1, T_2, A_1, A_2)$ выбирается то, которое дополнительно минимизирует функцию штрафа за неточные биекции $l_4(T_1, T_2, A_1, A_2)$:

$$
\begin{aligned}
l_4(T_1, T_2, A_1, A_2) \quad = \quad & \sum_{\substack{i,j \\ A_1(i)=A_2(j)\neq 0 \\ |A_1^{-1}(A_1(i))|=1 \\ |A_2^{-1}(A_2(j))|=1}} I\left(w^1_{A_1(i)} \neq w^2_{A_2(j)}\right) + \\
+ \quad & \sum_i I(A_1(i) = 0 \vee |A_2^{-1}(A_1(i))| > 1).
\end{aligned}
$$

Рассмотрим следующую функцию потерь:

$$\mathcal{L}_{strict}(T_1, T_2, A_1, A_2) = \sum_{i=1}^{3} \varepsilon_i l_i(T_1, T_2, A_1, A_2),$$

$$\mathcal{L}(T_1, T_2, A_1, A_2) = \mathcal{L}_{strict}(T_1, T_2, A_1, A_2) + \varepsilon_4 l_4(T_1, T_2, A_1, A_2),$$

где ε_1, ε_2, ε_3 и ε_4 — некоторые положительные коэффициенты.

Несложно видеть, что в связи с тем, что функции l_i принимают только целый значения, подбор коэффициентов ε_i позволяет минимизировать только функционал $\mathcal{L}(T_1, T_2, A_1, A_2)$ для достижения того же результата. Действительно, достаточно взять $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 1$ и $\varepsilon_4 = \frac{1}{M+1}$, где M — максимальное допустимое возможное количество токенов в тексте. При описанном наборе коэффициентов выполняется следующее соотношение.

$$\mathcal{L}(T_1, T_2, A_1, A_2) < 1 \Longleftrightarrow \mathcal{L}_{strict}(T_1, T_2, A_1, A_2) = 0$$

Задача нечёткого сопоставления пары токенизаций

Для данного случая был предложен алгоритм построения выравнивания, про который доказаны следующие теоремы.

Определение

Пара наборов слов T_1, T_2 является допустимой, если существует выравнивание, задаваемое A_1, A_2, такое что $\mathcal{L}_{strict}(T_1, T_2, A_1, A_2) = 0$.

Теорема

Для допустимой пары наборов слов T_1, T_2 описанный выше алгоритм строит оптимальное выравнивание A_1, A_2.

Теорема

Алгоритм имеет сложность $O(|T_1| \times |T_2|)$.

Список членов программного комитета

Задача извлечения членов программного комитета стоит из двух шагов:

- нахождение области, содержащий список членов программного;
- выделение в каждой строке полей имя, организация и страна.

Пример списка членов программного комитета:

```
Program Committee:
Sihem Amer-Yahia                    CNRS LIG, France
Gildas Avoine                            UCL, Belgium
Boualem Benatallah                  UNSW, Australia
Gregor V. Bochmann                Univ of Ottawa, Canada
```

Обнаружение областей с описанием персон

Результат применения внешнего модуля распознавания именованных сущностей:

```
Program Committee:
Sihem/PER Amer-Yahia/PER     CNRS LIG , France/LOC
Gildas/PER Avoine/PER                 UCL/ORG, Belgium/LOC
Boualem/PER Benatallah/PER      UNSW , Australia/LOC
Gregor/PER V/PER . Bochmann/PER   Univ of Ottawa/LOC , C
Ahmed/PER Bouajjani/PER         Univ/PER Paris/PER 7 , Fr
```

В отсутствии лексического контекста модуль распознавания именных сущностей.

Обнаружение областей с описанием персон

Добавим информацию из автоматически построенного по Википедии газитира университетов, стран и их псевдонимов.

```
Program Committee:
Sihem Amer-Yahia        CNRS LIG, France/COUNTRY
Gildas Avoine           UCL, Belgium/COUNTRY
Boualem Benatallah      UNSW/SCHOOL, Australia/COUNTRY
Gregor V. Bochmann      Univ/SCHOOL of/SCHOOL Ottawa/SCHOOL
Ahmed Bouajjani         Univ/SCHOOL Paris/SCHOOL 7/SCHOOL,
```

Информация из газитира дополняет данные модуля распознавания именных сущностей.

Обнаружение областей с описанием персон

Для обобщения двух разметок введём естественное понятие роли плитки: PER, ORG, LOC, O.
Отнесём к областям с описанием персон строки, которые:

- содержат все три плитки с ролями PER, ORG, LOC;
- содержат хотя бы одну плитку с ролью PER, ORG, LOC и находится рядом со строкой из области.

Обнаружение формата списка персон

Для отделения имени, организации и страны могут использоваться различные форматы:

```
Kiran-Kumar Muniswamy-Reddy, Amazon, USA.
Hamid Reza Motahari Nezhad, HP Labs, USA.
Peter Pietzuch, Imperial College, London, UK.
```

```
Duc A. Tran (UMass Boston, USA)
Maria da Gra\u00e7a Pimentel (Universidade de S\u00e3o P
Dominique Laurent (Universit\u00e9 de Cergy-Pontoise, Fr
```

```
 * Ling Chen        University of Technology, Sydney
 * Peter Dolog      Aalborg University, Denmark
 * Yan Li           University of Southern Queensland, Aus
 * Xue Li           University of Queensland, Australia
 * Chaoyi Pang      CSIRO, Australia
```

Обнаружение формата списка персон

- Данные об именованных сущностях недостаточно точны для выделения подполей
- Формат данных заранее не известен

Обнаружение формата списка персон

- Данные об именованных сущностях недостаточно точны для выделения подполей
- Формат данных заранее не известен

Идея

Используя частичную информацию об именованных сущностях определить формат записи подполей

Обнаружение формата списка персон

Для работы с форматом введём следующие обозначения:

- **Роль токена.** Если все плитки, покрывающие токен имеет одну и ту же роль, то она считается ролью токена; если токен не покрыт плитками, то его роль — 0; в противном случае роль не определена.

```
Gildas/PER  Avoine/PER                    UCL/ORG, Belgium/LOC
```

Обнаружение формата списка персон

Для работы с форматом введём следующие обозначения:

- **Роль токена.**
- **Ролевой шаблон строки.** Заменим в строке все токены-слова на их роли, все токены пунктуации на их текстовое представления, длинные пробелы на специальный символ `long_space` и удалим все прочие токены. Полученная после удаления повторяющихся элементов последовательность называется ролевым шаблоном строки.

```
Gildas/PER Avoine/PER                    UCL/ORG, Belgium/LOC
PER long_space ORG , LOC
```

Обнаружение формата списка персон

Для работы с форматом введём следующие обозначения:

- **Роль токена.**
- **Ролевой шаблон строки.**
- **Соответствие ролевого шаблона строке.** Будем говорить, что строка соответствует ролевому шаблону, если каждому элементу шаблона можно поставить в соответствие непрерывный сегмент строки, причём знаки пунктуации и `long_space` отображаются в ровно один идентичный токен.

```
Gildas/PER Avoine/PER                UCL/ORG , Belgium/LOC
PER long_space ORG , LOC
PER long_space LOC , ORG
PER long_space PER , ORG
PER , ORG , LOC
```

Обнаружение формата списка персон

Для работы с форматом введём следующие обозначения:

- **Роль токена.**
- **Ролевой шаблон строки.**
- **Соответствие ролевого шаблона строке.**
- **Соответствие шаблона и данных**: как выбрать лучший шаблон?

```
Gildas/PER Avoine/PER                UCL/ORG, Belgium/LOC
PER long_space ORG , LOC
PER long_space LOC , ORG
PER long_space PER , ORG
PER , ORG , LOC
```

Вероятностная модель порождения ролевого шаблона

Необходимо оценить вероятность порождения шаблона P по данным D и выбрать лучший шаблон $\hat{P}$:

$$\hat{P} = \arg\max_{P} \mathrm{P}(P|D)$$

Вероятностная модель порождения ролевого шаблона

Необходимо оценить вероятность порождения шаблона P по данным D и выбрать лучший шаблон $\hat{P}$:

$$\hat{P} = \arg\max_P \mathrm{P}(P|D) = \arg\max_P \mathrm{P}(P)\,\mathrm{P}(D|P)$$

Вероятностная модель порождения ролевого шаблона

Необходимо оценить вероятность порождения шаблона P по данным D и выбрать лучший шаблон $\hat{P}$:

$$\hat{P} = \arg\max_P \mathrm{P}(P|D) = \arg\max_P \mathrm{P}(P)\,\mathrm{P}(D|P)$$

Данные - это набор строк, где каждая строка представлена парой (токены, плитки): $D = \{(W_1, T_1), \dots\}$.

$$\hat{P} = \arg\max_P \mathrm{P}(P)\,\mathrm{P}(D|P) := \arg\max_P \mathrm{P}(P) \prod_{i=1}^{|D|} \mathrm{P}(W_i, T_i|P)$$

- $\mathrm{P}(P)$ — априорная вероятность шаблона P;
- $\mathrm{P}(W, T|P)$ — правдоподобие строки (W, T) при заданном шаблоне P.

Вероятностная модель порождения ролевого шаблона

Как определить $\mathrm{P}(W, T|P)$, где

- $W = (w_1, \ldots, w_{|W|})$ — набор токенов;
- $T = (t_1, \ldots, t_{|T|})$ — набор плиток;
- $P = (p_1, \ldots, p_{|P|})$ — ролевой шаблон?

Сложность:

- последовательности разной длины;
- бесконечное количество слов.

Вероятностная модель порождения ролевого шаблона

Как определить $P(W, T|P)$?
Сложность:

- последовательности разной длины;
- бесконечное количество слов.

Решение:

- введём скрытую переменную S, отвечающую за соответствие шаблона и токенов;
- выполним факторизацию с помощью подхода, основанном на графических моделях.

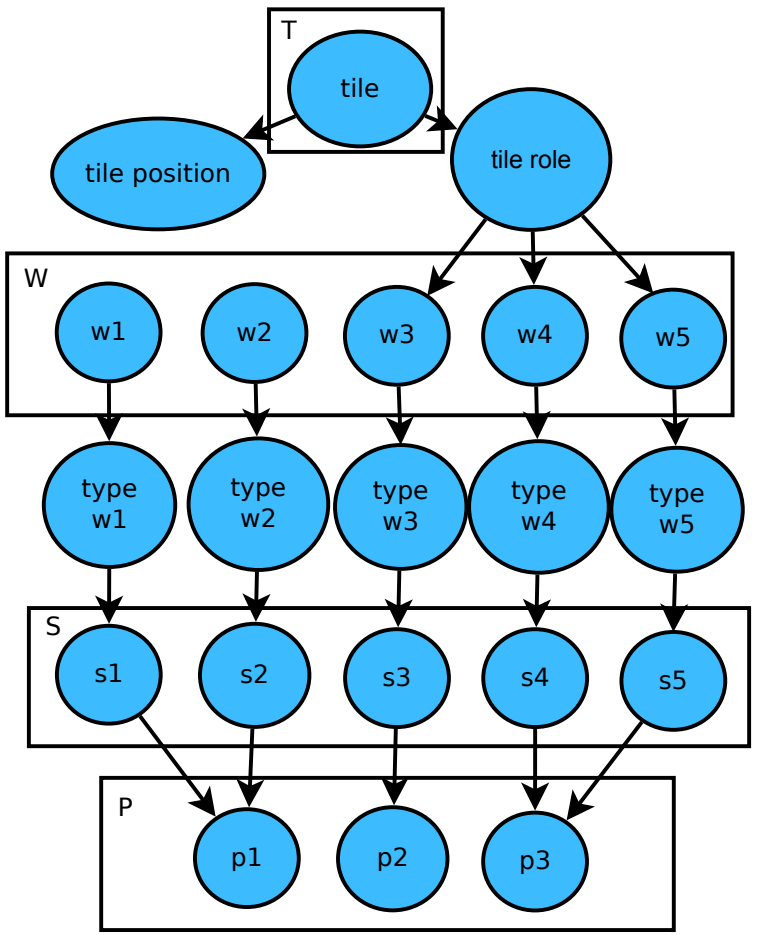
Вероятностная модель порождения ролевого шаблона

Как определить $P(W, T|P)$?
Сложность:

- последовательности разной длины;
- бесконечное количество слов.

Решение:

- введём скрытую переменную S, отвечающую за соответствие шаблона и токенов;
- выполним факторизацию с помощью подхода, основанном на графических моделях.



Определение (Роль токена)

Рассмотрим все плитки, накрывающие данный токен. Если не существует ни одной накрывающей плитки, или все плитки имеют тривиальные роли, то положим по определению, что токен имеет роль other. Если все накрывающие плитки с нетривиальными ролями имеют одинаковую роль, то она присваивается и токену. В противном случае роль токена не определена. Будем обозначать роль токена t за $\rho(t)$.

Определение

Пусть $\mathfrak{M}$ — модель текста. Ролевой шаблон — конечная последовательность P элементов множества $\{*, person, organization, location\} \cup \Sigma^+ \cup \{longspace\}$, где Σ — алфавит модели $\mathfrak{M}$. Будем называть элемент этой последовательности нетривиальным, если он принадлежит множеству $\{person, organization, location\}$.

Определение

Отображением ролевого шаблона P на последовательность токенов T называется такое отображение $S : \{1, \ldots, |P|\} \to \mathfrak{U}(\{1, 2, \ldots, |T|\})$, где $\mathfrak{U}$ — множество всех подмножеств, для которого выполнены следующие условия:

- $\forall i : S(i) \neq \emptyset$;
- $\forall i \, \forall j > i \, \forall p_i \in S(i) \, \forall p_j \in S(j) : p_i < p_j$;
- $\forall p \in \{1, 2, \ldots, |T|\} \, \exists i : p \in S(i)$.

Определение (Соответствие последовательности токенов ролевому шаблону)

Будем говорить, что непустая последовательность токенов $T = \{t_1, \ldots, t_{|T|}\}$ *удовлетворяет непустому ролевому шаблону* $P = \{p_1, p_2, \ldots, p_{|P|}\}$, *если существует такое отображение* S *шаблона* P *на последовательность токенов* T, *что выполнены следующие условия:*

$$\begin{aligned} \forall i : p_i \in \Sigma^+ \quad &\Rightarrow \quad |S(i)| = 1 \wedge (\forall s \in S(i) : text(t_s) = p_i), \\ \forall i : p_i = longspace \quad &\Rightarrow \quad |S(i)| = 1 \wedge (\forall s \in S(i) : \\ &\qquad type(t_j) = space \ \wedge \ islongspace(t_s)), \end{aligned}$$

где $text(t)$ *— текст токена* t, $type(t)$ *— тип токена* t, *а предикат* $islongspace(t)$ *принимает значение истина тогда и только тогда когда* $text(t)$ *либо содержит более одного пробельного символа, либо содержит символ табуляции. Такое отображение будет называть корректными и обозначать множество таких отображений за* $P \models T$.

Алгоритм построения списка шаблонов для строки:

- Вход:
  - список токенов T.
- Выход:
  - ролевой шаблон P.
- Инициализация:
  - создадим пустой ролевой шаблон P.
- Для каждого элемента t из T:
  - если роль токена определена, то положим $\rho(t)$ в P;
  - иначе если тип токена — *punct*, то положим текст данного токена в P;
  - иначе если истинно $islongspace(t)$, то положим *longspace* в P;
  - иначе если тип токена — *space* или роль токена не определена, то пропустим данный токен;
  - иначе положим $*$ в P.
- Удалим идущие подряд повторяющиеся элементы из P.

Блок D состоит из k строк, каждая строка суть пара (W_i, T_i), где W_i — последовательность токенов $w_{i,1}, w_{i,2}, \ldots, w_{i,|W_i|}$, а T_i — множество плиток $\{t_{i,1}, \ldots, t_{i,|T_i|}\}$.
Условие оптимальности шаблона:

$$\hat{P} = \arg\max_{P \in \mathbb{P}} \mathrm{P}(P|D) = \arg\max_{P \in \mathbb{P}} Score(P),$$

$$Score(P) = \Bigg(\prod_{i=1}^{k} \sum_{S \in (P \models W_i)} \Big(\prod_{j=1}^{|T_i|} \mathrm{P}(t_{i,j}^{role} | P_{S^{-1}(t_{i,j})}) \prod_{j=1}^{|W_i|} \mathrm{P}(w_{i,j}^{type} | p_{S^{-1}(j)}) \, \mathrm{P}(S|P) \Big) \Bigg) \mathrm{P}(P_*) \, \mathrm{P}(P_{uniq\_cnt}) \, \mathrm{P}(P_{duplicate}) \Bigg).$$

Обнаружение формата списка персон

В введённых обозначениях для рассматриваемого примера алгоритмом будут построены следующие роли:

```
Program Committee:
Sihem/PER Amer-Yahia/PER      CNRS/O LIG/O, France/LOC
Gildas/PER Avoine/PER         UCL/ORG, Belgium/LOC
Boualem/PER Benatallah/PER    UNSW/ORG, Australia/LOC
Gregor/PER  V/PER . Bochmann/PER   Univ/ORG of/ORG Ottawa/UNDEF, Canada/LOC
Ahmed/PER Bouajjani/PER       Univ/ORG Paris/SCHOOL 7/SCHOOL, France/LOC
```

Далее по каждой строке строится ролевой шаблон:

- `PER long_space * , LOC`
- `PER long_space ORG , LOC`
- `PER long_space ORG , LOC`
- `PER . PER long_space , LOC`
- `PER long_space ORG , LOC`

Обнаружение формата списка персон

Значения функции ценности для данного примера:

- P(person long_space * , location | D) $\propto 5.0 \times 10^{-32}$;
- P(person long_space organization , location) $\propto 9.5 \times 10^{-31}$;
- P(person . person long_space organization , location) $\propto 0$.

Легко видеть, что наибольший вес получит правильный вариант, который и будет применён ко всем элементам списка.

Общий план

Алгоритмы анализа метаинформации о научно-технических конференциях

Наибольший интерес в рамках задачи анализа динамики направлений научных исследований представляет:

- тематика конференции;
- страна проведения конференции;
- временны́е данные, такие как дата проведения конференции.

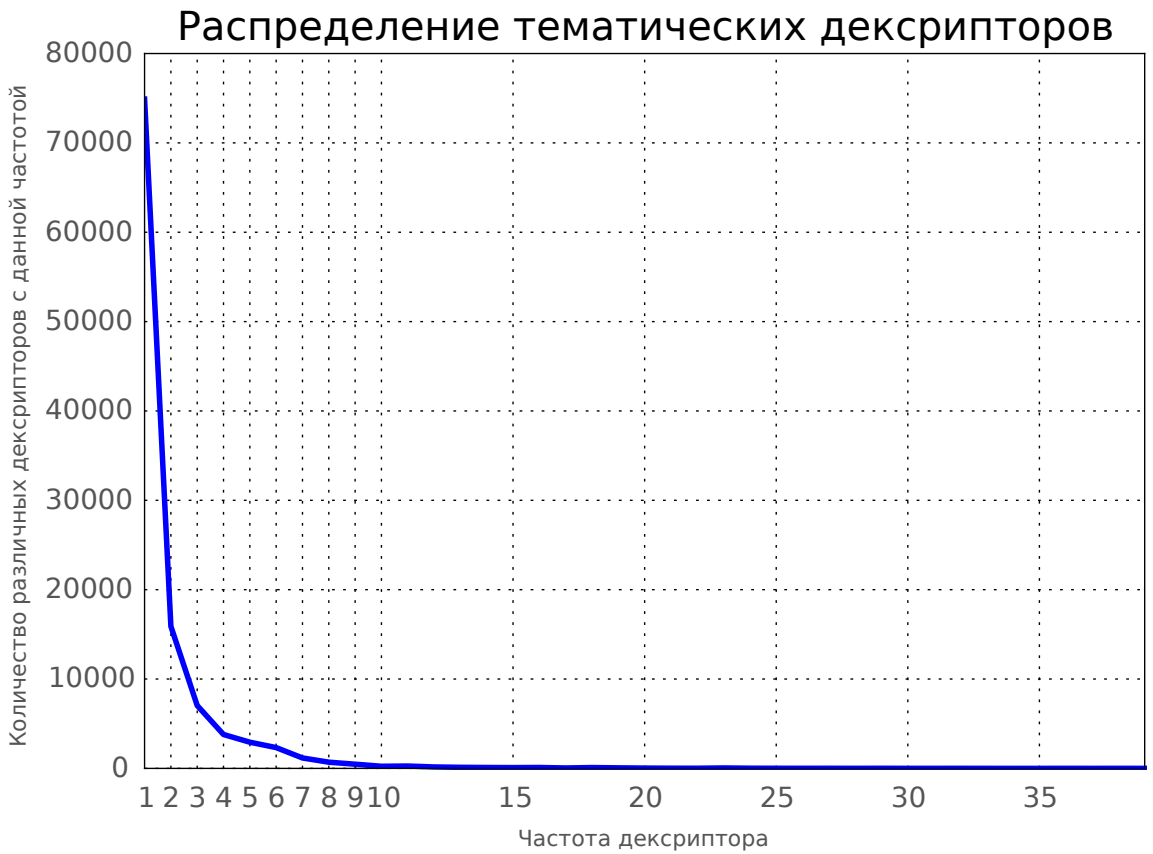
Необходимо также придать данным семантику, т.е. иметь возможность отвечать на вопросы:

- относится ли конференция к определённой области?
- в каком регионе находится данные страна?
- в каком году/месяце произошла конференция?

Анализ тематических дескрипторов

Наивный подход: объявить каждый дескриптор отдельным направлением исследования.

Данный подход не работает из-за проблем с разреженностью.



Анализ тематических дескрипторов

Причины разреженности тематических дескрипторов:

- специализация;
- синонимия в естественном языке.

Пример списка ключевых тематик с конференции WiSec 2014:

- Mobile malware and platform security
- Security & Privacy for Smart Devices (e.g., Smartphones)
- Wireless and mobile privacy and anonymity
- Secure localization and location privacy
- Cellular network fraud and security
- Jamming attacks and defenses

Анализ тематических дескрипторов

Необходимо проверить принадлежность тематического дескриптора некоторой области.

Можно использовать составленную вручную классификаторы (Универсальная десятичная классификация (УДК), классификация ассоциации вычислительной техники (ACM Computing Classification System))

Недостатки классификаторов:

- не решает проблемы множества написаний;
- консервативность;
- недостаточная подробность.

Анализ тематических дескрипторов

Необходимо проверить принадлежность тематического дескриптора некоторой области.

Можно строить иерархию понятий автоматически.

Построение онтологии, которая охватывает все возможные области науки, может занять огромное количество времени и ресурсов.

Однако достаточно иметь возможность отвечать на следующие запросы:

- получить список вышестоящих понятий (поиск гиперонимов);
- получить список нижестоящих понятий (поиск гипонимов);
- получить список аналогичных понятий (поиск когипонимов).

Лексические шаблоны

Гипотеза (Hearst, 1992)

Пусть T - гипоним (частное), H - гипероним (общее), тогда можно ожидать, что они будут встречаться в шаблонах вида:

T is a H | *T is a kind of H*
H including T | *H such as T*

Предполагается, что и обратное утверждение верно.

В оригинальной работе предлагается алгоритм поиска гипонимов, основанный на применении шаблона «H such as *» и использующий в качестве источника информации энциклопедию.

Энциклопедии: (+) синтаксически корректные тексты;
(–) ограниченный терминологический охват, медленное обновление

В рамках данной работы в качестве источника использовались тексты из сети Интернет.

Применение метода Hearst в сети Интернет

Для нахождения гиперонимов, например, слова *chimpanzee* можно выполнить к поисковой интернет-системе запрос
`"chimpanzee is a"`[1],
и выбрать слова, стоящие справа от «is a».
Ниже приведены найденные гиперонимы для слов *chimpanzee* (шимпанзе) и *orthodoxy* (православие). В скобках — количество найденных фрагментов с данным гиперонимом.

- Для слова *chimpanzee* : **ape** (59), **animal** (34), hug (22), bust (18), de (16), model (15), species (12), . . . (всего 138 вариантов)
- Для слова *orthodoxy*: series (26), book (20), movement (17), way (13), part (10), wave (9), **religion** (9), . . . (всего 180 вариантов)

Проблемы:

- большое число вариантов возможных гиперонимов
- частотность не является критерием корректности

[1] Кавычки заставляют поисковую систему искать точное вхождение

Проблемы обработки текстов из Интернета

Ниже приведены фрагменты, возвращаемые поисковой системой, иллюстрирующие основные проблемы, связанные с использованием лексических шаблонов.

- Chimpanzee is a monkey that relatively strong with a small body size
- AK47 is a great gun for people looking to break out of starter airsoft guns
- Why tequila is a girl's best friend.
- Elementary Algebra is a branch of mathematics of generalized ...
- In a nutshell, Ubuntu is a Linux distribution for people.
- Send-Safe Proxy Scanner is a program designed for searching for HTTPS/SOCKS ...

Этапы предлагаемого алгоритма

Этапы предлагаемого алгоритма:

- С помощью шаблона «T is a» извлечь из Интернета слова — кандидаты в гиперонимы для данного слова
- Для «обрубленных» кандидатов найти более точные, расширенные варианты — словосочетания
- Отфильтровать полученное множество множество пар (гипоним, предполагаемый гипероним).

Фильтрация результатов

Для пары (T,H) необходимо определить, является ли она корректной гипонимической парой.

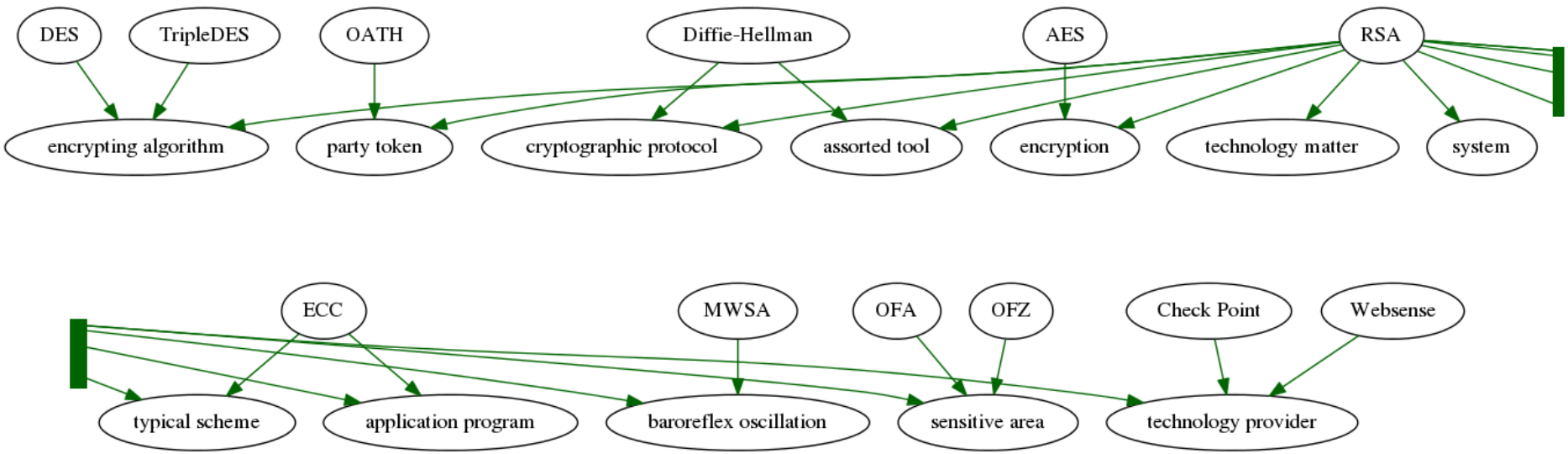
Можно вычислить следующие числовые признаки:

- Количество страниц, возвращаемых поисковой системой, для специальных запросов:
  - Левые шаблоны (например, «T is a H»)
  - Правые шаблоны (например, «H such as T»)
  - Нормирующие шаблоны (например, «T» и «T is a»)
- Количество братьев T (т.е. слов с общим гиперонимом H)

Братья извлекаются из фрагментов вида «`special attention must be paid to the renovationist currents in orthodoxy, islam, lamaism and other religions`».

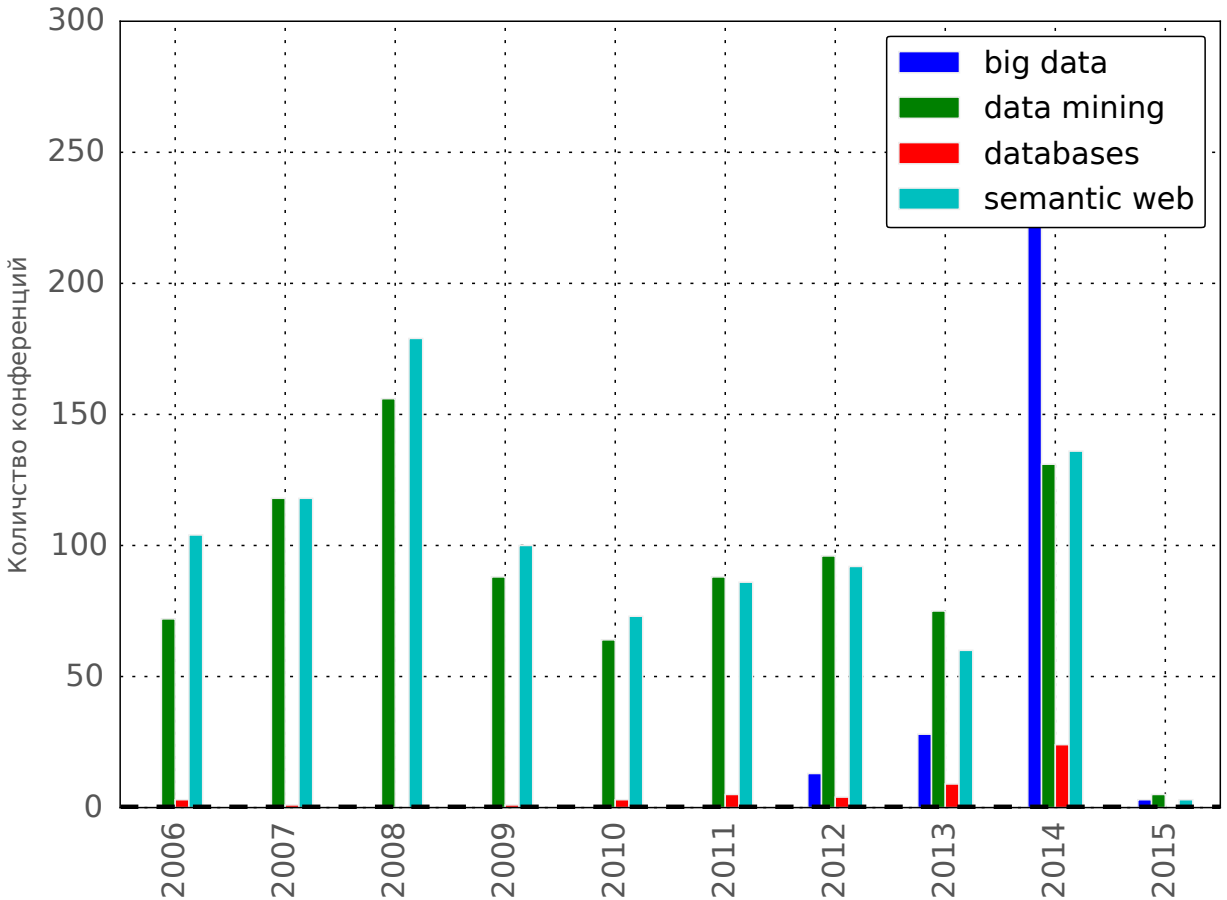
Анализ динамики актуальности научных областей

Пример применения алгоритма поиск гиперонимов к термину **RSA**.



Анализ динамики актуальности научных областей

Количество конференций по выбранным темам, найденных по точному вхождению.



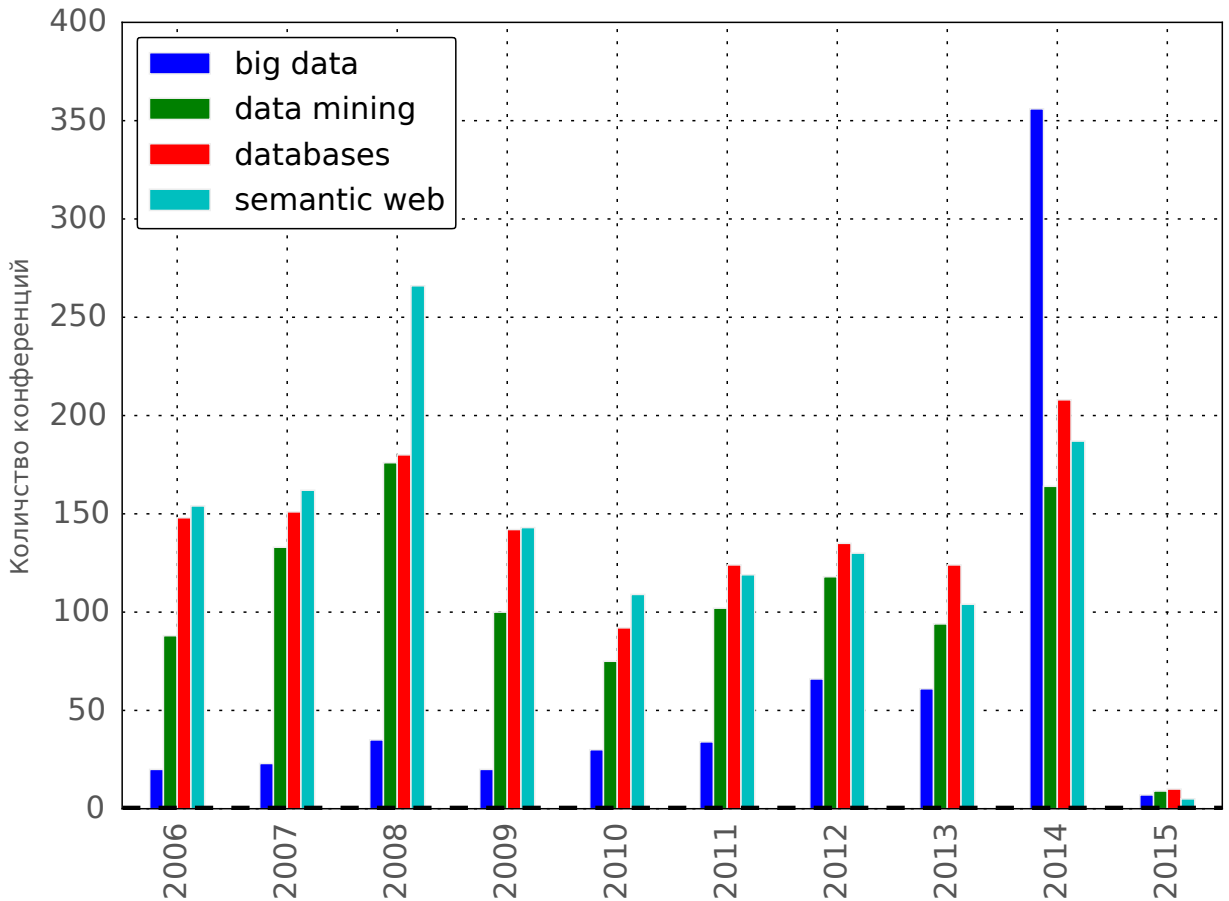
Анализ динамики актуальности научных областей

Найденные алгоритмом частные понятия для выбранных областей.

big data	hadoop, vertica, apache spark, nosql, data security solutions for hadoop, earth-level simulations, recommendation systems, big data, a/b testing, bioinformatics, hive, social feeds, pig, analytics, sas, hbase, credit histories, insurance, cloud, cassandra, genomics, web analytics, java, scala, storage, administrative data, financial reports, mccoy, ad preferences, text mining, python/pandas, finance, python, mongodb, nosql stores, seismic processing, internet data, microsoft dynamics, mahout [mah13, video surveillance, 3d scene reconstruction, salesforce, meteorology, mapreduce, spss, search history, r, parquet, deep analytics, com, telematics
databases	etags, quickbooks, databases, gene ontology, sms, sap, electronic health records, excel, access, lexus nexis, mms, refworks, ebscohost, sisense, oracle, wikipedia, www, ms sql server, contact managers, electronic medical records, spreadsheets, westlaw, mysql, factset, email / accounts programs, citation databases, manupatra, log itself, erp systems, commercial, uniprotkb, scientific search engines, amadeus, sql, siebel, proquest 's historical newspapers, strain collections, ms office, morningstar, cia world factbook, heinonline, zip code, websites, cochrane, lexis nexis, raiser 's edge, dxcc, area code, public records
semantic web	owl, plata, sparql, semantic web, rdf, wordnet, rda tags, metadata, inferencers for rdf, multimedia files, web-pages, xliff, its
data mining	data mining, neural networks, sas, google trends, intelligent software agents, biorepository, standard deviation, pipeline maintenance, cluster analysis, mean, smart tags, text, keystroke loggers, boolean search methodology, financial analytics, acl, dss, comfa, toad, rfid, enantiophores/pharmacophore modelling, medical imaging, sql, linear regression, qsprs, olap, foot-fall counts, comparative collection assessment project

Анализ динамики актуальности научных областей

Количество конференций по выбранным темам, найденных с использованием гипонимов.



Общий план

Компоненты системы

Для проведения аналитических экспериментов и апробации описанных алгоритмов был реализован прототип системы сбора и анализа сообщений о научных конференциях.

Основные функции:

- загрузка сообщений из сети Интернет;
- применение правил к сообщениям;
- поиск и визуализация сообщений;
- редактирование эталонной разметки сообщений;
- построение гипонимов и гиперонимов.

Оценка качества системы извлечения данных

Качество извлечения данных измерялось с помощью сравнения с эталонным извлечением, выполненным вручную.

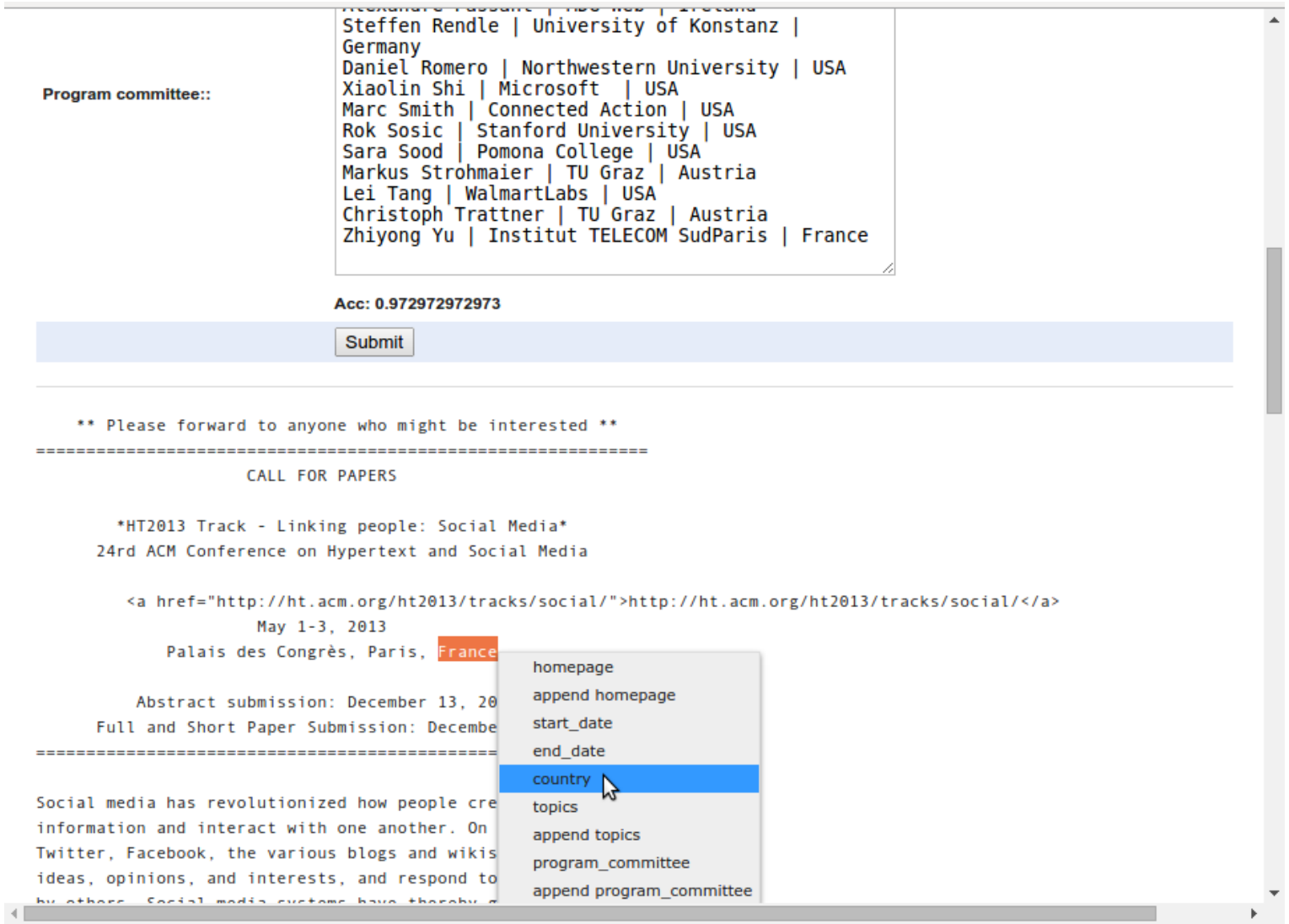
Численная оценка соответствия результата работы алгоритма эталонному производилась с помощью коэффициент Сёренсена, который для двух конечных множеств A и B вычисляется по формуле:

$$K(A, B) = \frac{2|A \cap B|}{(|A| + |B|)}$$

.

Оценка качества системы извлечения данных

Фрагмент интерфейса редактирования эталонной разметки.



Оценка качества системы извлечения данных

Значение коэффициента Сёренсена на контрольной выборке:

Название поля фрейма	**Точность в процентах**
Дата начала конференции	100.00%
Дата завершения конференции	100.00%
Страна проведения конференции	100.00%
Домашняя страница конференции	100.00%
Список членов программного комитета	88.16%
Ключевые тематики конференции	96.91%

Общий план

Основные результаты

- Разработана формальная модель хранения текста и метаинформации из текстов сообщений о научно-технических конференциях. Данная модель позволяет извлекать произвольное количество полей с учётом графической разметки и других особенностей неструктурированных текстов.
- Предложен эффективный алгоритм построения выравнивания между двумя искажёнными токенизациями одного текста. Доказана его корректность и оценка вычислительной сложности.
- Разработан алгоритм построения иерархии научных понятий на основе анализа текстов из сети Интернет и его программная реализация. Продемонстрирована эффективность использования данного алгоритма в рамках решения задачи анализа актуальности научных направлений.
- Создан прототип системы сбора сообщений о научно-технических конференциях и извлечения метаинформации из них. Проведены тестовые испытания системы, продемонстрировавшие её работоспособность и высокую точность извлечения метаинформации.

Спасибо за внимание!

Структура диссертации

Диссертация состоит из четырёх глав.

1. Анализ методов исследования активности научного сообщества.
2. Алгоритмы извлечения метаинформации из сообщений о научно-технических конференциях.
3. Алгоритмы анализа метаинформации о научно-технических конференциях.
4. Программная реализация.

Публикации

Опубликовано 5 работ (из них 3 в списке ВАК)

- *Бахтин А.В.* К созданию программного комплекса для извлечения метаинформации о научно-технических конференциях // Программная инженерия. — 2013. — № 11. — С. 32–38
- *Васенин В.А., Афонин С.А., Козицын А.С., Бахтин А.В.* Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) // Обозрение прикладной и промышленной математики. — 2012. — Т. 19, № 2. — С. 239–240.
- *Афонин С.А, Бахтин А.В.* Построение иерархии понятий на основе лексических шаблонов // Информационные технологии. — 2012. — № 3. — С. 2–7.
- *Афонин С.А., Бахтин А.В.* Об одном методе построения гипернимов с помощью внешней поисковой системы // Материалы Всероссийской конференции с международным участием «Знания - Онтологии - Теории» (ЗОНТ-2011), 3-5 октября 2011 г., Новосибирск. — Т. 1. — Институт математики им. С.Л. Соболева СО РАН, Новосибирск, 2011. — С. 14–24
- *В.А. Садовничий, С.А. Афонин, А.В. Бахтин, В.Ю. Бухонов, В.А. Васенин, Г.М. Ганкин, А.Э. Гаспарянц, Д.Д. Геломазов, А.А. Иткес, А.С. Козицын, И.Н. Тумайкин, К.А. Шапченко.* Интеллектуальная система тематического исследования научно-технической информации («ИСТИНА»). Издательство Московского университета Москва, 2014.

Расширение понятий

Причины выделения неправильных кандидатов — либо в исходном фрагменте определялось что-то не то, что мы искали, либо не так, как мы ожидали. Рассмотрим пример: *Send-Safe Proxy* ***Scanner*** *is a* ***program*** *designed for searching for HTTPS/SOCKS . . .*
Выполним запрос: «* scanner is a program»
Полученные значения для звёздочки: proxy(147), a(56), ftp(51), port(29), . . .
Имеется явный лидер, поэтому можно считать, что значение *scanner*, к которому относится гипероним program, — побочное.

Рассмотрим пример: *In a nutshell,* ***Ubuntu*** *is a Linux* ***distribution*** *for people.*
Выполним запрос: «Ubuntu is a * distribution»
Полученные значения для звёздочки: Linux(40), good(3), . . .
Имеется явный лидер, которого стоит рассматривать в качестве альтернативного гиперонима.