Метрики семантической близости с приложениями к задачам АОТ

Александр Панченко
Université catholique de Louvain
alexander.panchenko@uclouvain.be

4 апреля 2013 г.

План

План

Введение

Мотивация

1. Метрики семантической близости полезны для:
   - систем обработки коротких текстов (Šaric et al., 2012; Panchenko at., 2012);
   - расширешия поисковых запросов (Hsu et al., 2006);
   - вопросно-ответных систем (Sun et al., 2005);
   - разрешения омонимии (Patwardhan et al., 2003);
   - ...

- Лексико-семантическое знание о языке.
- Вычислительная лексическая семантика.
- Computational Lexical Semantics.

Метрики семантической близости

Определение

Метрика семантической близости численно выражает семантическую связность двух c_i, c_j: $s_{ij} = sim(c_i, c_j)$:

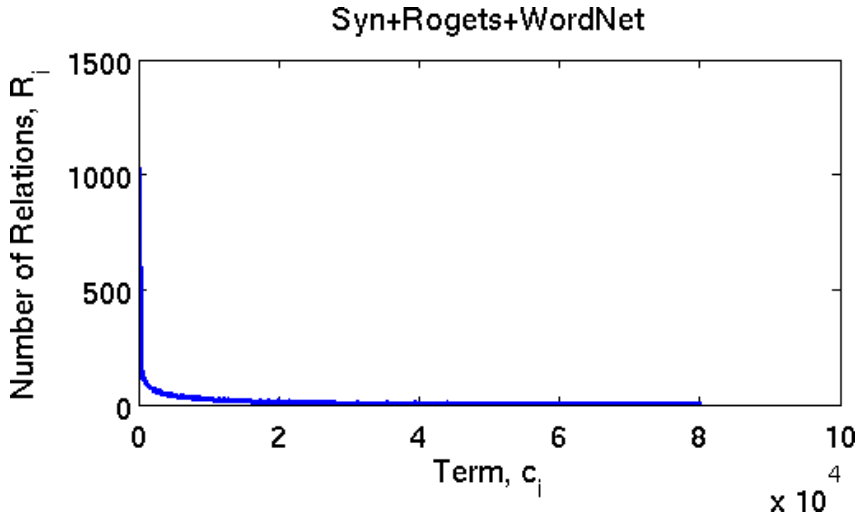
$$s_{ij} = \begin{cases} \text{велико} & \text{если } \langle c_i, c_j \rangle - \text{пара } syn, hyper, cohypo \\ 0 & \text{иначе} \end{cases}$$

Свойства

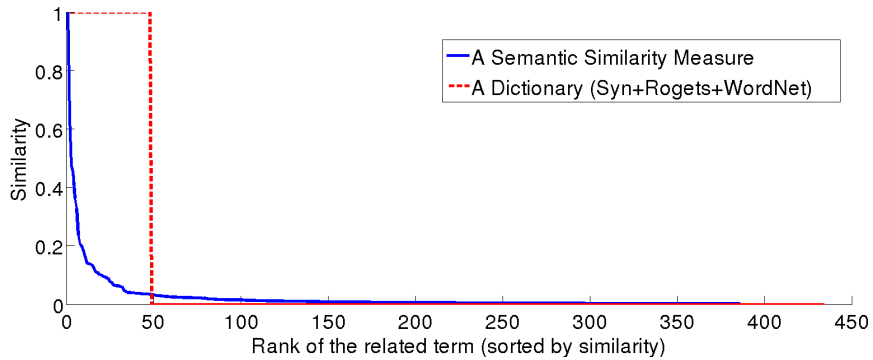
- Неотрицательность: $0 \leq s_{ij} \leq 1$;
- Рефлективность: $s_{ij} = 1 \Leftrightarrow c_i = c_j$;
- Симметричность: $s_{ij} = s_{ji}$;
- $s_{ij} \leq s_{ik} + s_{kj}$

Метрики семантической близости

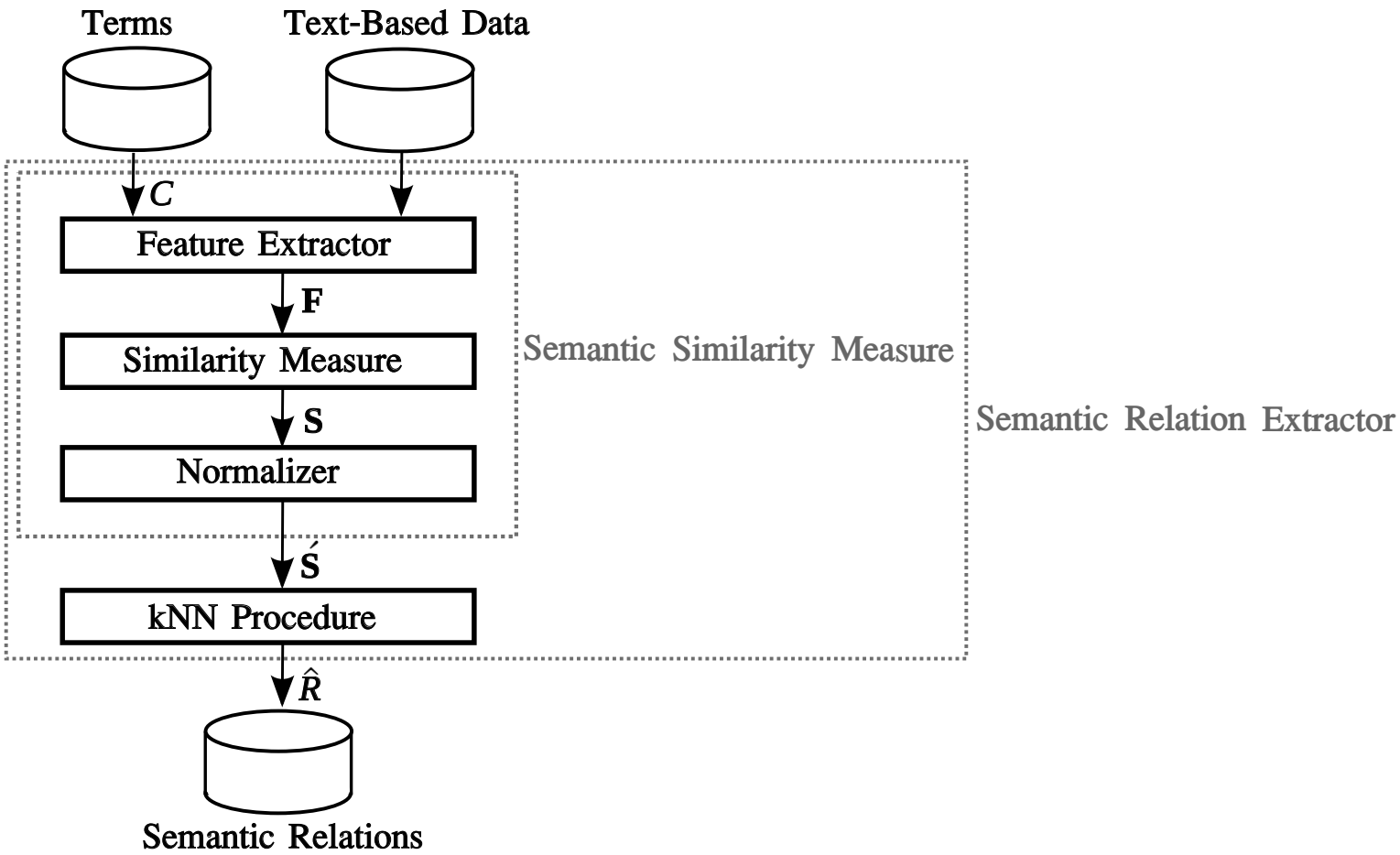
- Малое количество подобных пар: $s_{ij} \sim exp(\lambda)$.



- Распределение сем. близости слова “doctor”:



Системы измерения семантической близости



Как построить систему с высокой **точностью** и **лексическим покрытием**?

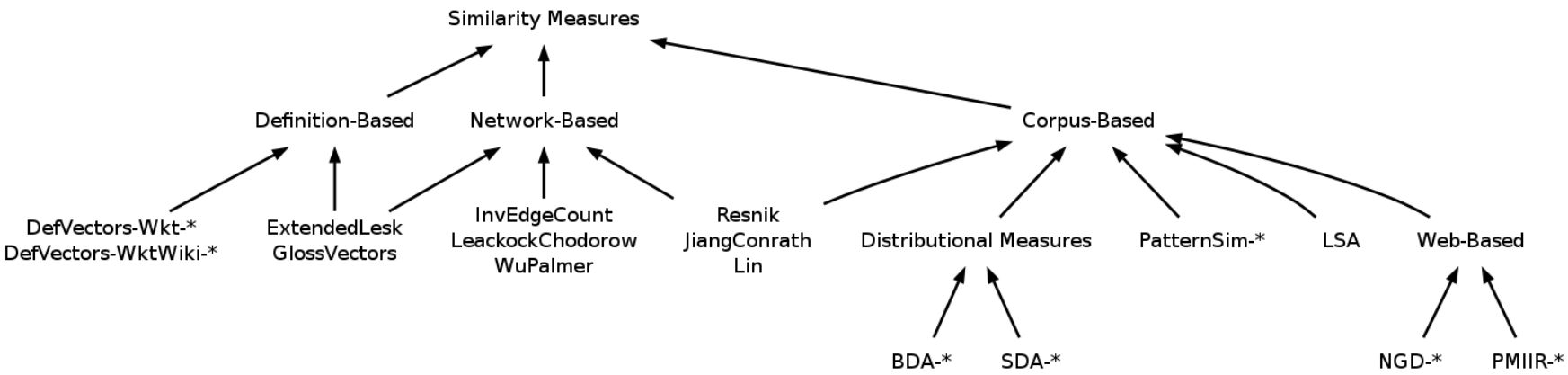
Оценка качества метрики семантической близости

1. **Корреляция с суждениями человека о сем. близости**:
   - Статистики: корреляция Пирсона (ρ) и Спирмена (r).
   - Проверочные данные: MC, RG, WordSim.
2. **Ранжирование семантических отношений**:
   - Точность, Полнота, F-мера.
   - Проверочные данные: BLESS, SN.
3. **Точность извлечения семантических отношений:**
   - Статистики: Точность@k.
   - Проверочные данные: аннотирование и/или тезаурусы.
4. **Использование метрики в системе AOT:**
   - в системе классификации имен файлов (**iCOP**);
   - с системе поиска семантически связанных слов (**Serelex**).

Panchenko A., **Similarity Measures for Semantic Relation Extraction.** PhD thesis. Université catholique de Louvain. 197 pages, 2013, (Chapter 1).

План

Обзор метрик семантической близости



Публикации (анализ 37 базовых метрик):

- Panchenko A., **Similarity Measures for Semantic Relation Extraction.** PhD thesis. Université catholique de Louvain. 197 pages, 2013, (Chapter 3).
- Panchenko A. **A Study of Heterogeneous Similarity Measures for Semantic Relation Extraction.** // In JEP-TALN-RECITAL 2012 — Grenoble (France), 2012.

Метрики основанные на семантической сети

Данные: семантическая сеть WordNet 3.0, корпус SemCor.
Переменные:

- $len(c_i, c_j)$ – длина **кратчайшего пути** между c_i и c_j
- $len(c_i, lcs(c_i, c_j))$ – длина кратчайшего пути от c_i до **ближайшего общего предка (БОП)** слов c_i и c_j
- $len(c_{root}, lcs(c_i, c_j))$ – длина кратчайшего пути от **корня** c_{root} до БОП слов c_i и c_j (глубина БОП)
- $P(c)$ – **вероятность слова** c, оцененная из корпуса
- $P(lcs(c_i, c_j))$ – **вероятность БОП** слов c_i и c_j

Метрики: Инвертированная длина пути (Jurafsky and Martin, 2009), Leacock-Chodorow (1998), Wu-Palmer (1994), Resnik (1995), Jiang-Conrath (1997), Lin (1998).

Метрики основанные на Веб корпусе текстов

Данные: количество документов возвращенных ИПС: Google, Yahoo, AltaVista, Bing, и т.п.

Переменные:

- h_i – **количество документов** возвращенных по запросу слова "c_i"
- h_{ij} – **количество документов** возвращенных по запросу "c_i AND c_j"

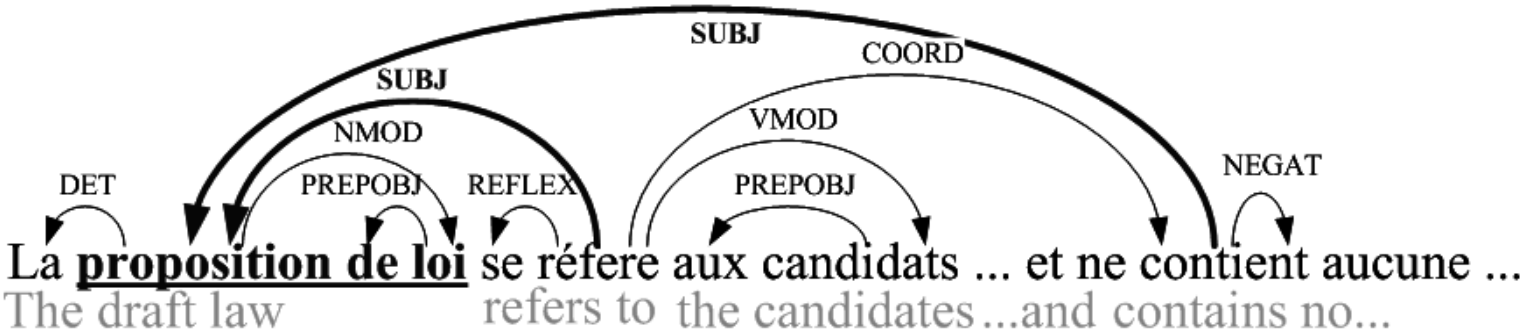
Метрики:

- Normalized Google Distance (NGD) (Cilibrasi and Vitanyi, 2007)
- Pointwise Mutual Information - Information Retrieval (PMI-IR) (Turney, 2001)

Дистрибутивные метрики

Данные: корпус, такой как Википедия или ukWaC
Переменные:

- $\mathbf{f}_i$– вектор признаков представляющий слово c_i, основанный на **контекстном окне**
- $\mathbf{f}_i^s$ – вектор признаков представляющий слово c_i, основанный на **синтаксическом контекстном окне**



Метрики:

- Bag-of-words Distributional Analysis (BDA) (Sahlgren, 2006)
- Syntactic Distributional Analysis (SDA) (Curran, 2003)

Другие метрики основанные на корпусе текстов

Данные: корпус, такой как Википедия или ukWaC
Метрики:

- Латентно-семантический анализ (LSA) (Landauer and Dumais, 1997)
- Вероятностные модели (pLSA, LDA и др.) (Griffiths et al., 2007)
- NGD и PMI-IR (Veksler et al., 2008)
- ...

Метрики основанные на определениях

Данные: определения из WordNet, Википедии, Викисловаря или любого другого словаря.

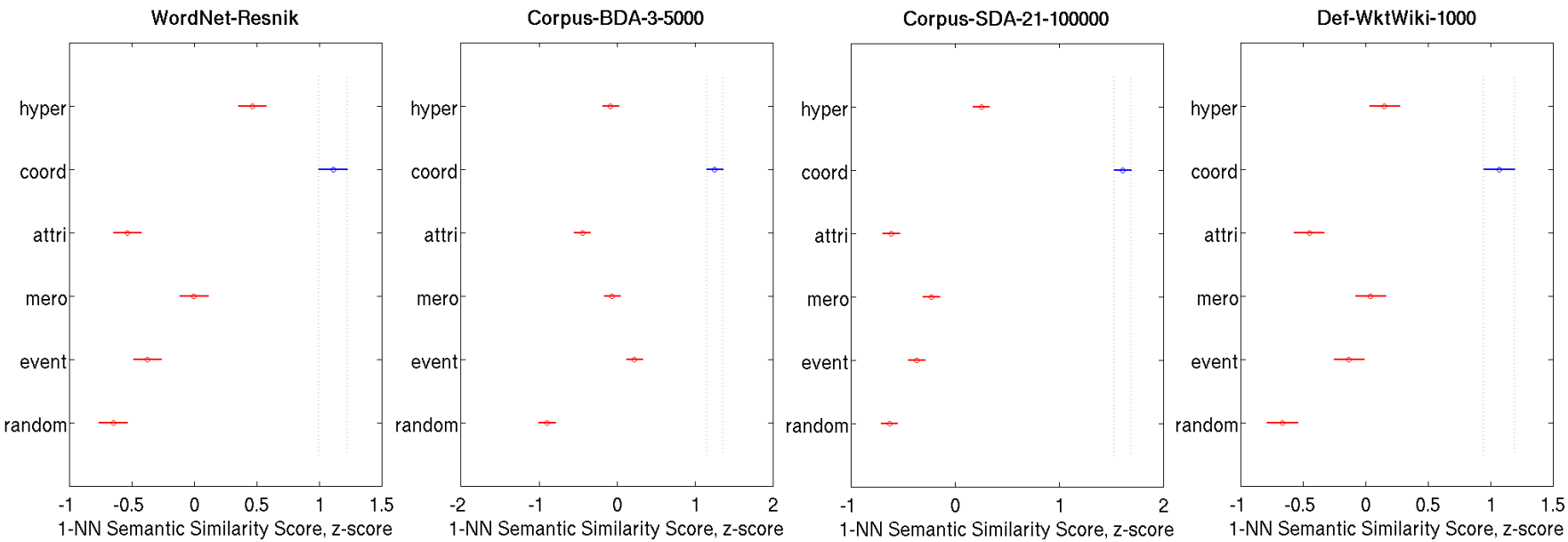
Переменные:

- $gloss(c_i)$ – **определение** слова c_i;
- $\mathbf{f}_i$ **вектор признаков**, построенный из $gloss(c_i)$;
- $\mathbf{f}_i$ – **вектор признаков** c_i, вычисленный на корпусе из всех определений методом контекстного окна;
- $exist(c_i, c_j)$ – наличие связи между c_i и c_j в словаре.

Метрики:

- ExtendedLesk (Banerjee and Pedersen, 2003)
- GlossVectors (Patwardhan and Pedersen, 2006)
- DefVectors (Панченко и др., 2012), (Panchenko et al., 2012)

Лучшие базовые метрики семантической близости



- Каждая метрика излекает много ко-гипонимов:
  - $\langle Canon, Nikon \rangle$,
  - $\langle Lamborghini, Ferrari \rangle$,
  - $\langle Obama, Romney \rangle$.

Резюме

Основные ресурсы для построения метрик:

- семантические сети и тезаурусы;
- корпуса текстов;
- Веб корпус текстов;
- определения из словарей и энциклопедий.

Метрики **дополняют друг друга** в терминах:

- лексического покрытия;
- точности;
- типов извлекаемых отношений.

Программное обеспечение

- **Semantic Vectors:** `https://code.google.com/p/semanticvectors/`
- **S-Space Package:** `https://code.google.com/p/airhead-research/`
- **WordNet::Similarity:** `http://wn-similarity.sourceforge.net`
- **NLTK:** `http://nltk.googlecode.com/svn/trunk/doc/howto/wordnet.html`
- **WikiRelate!**
- **PatternSim / Serelex:** `http://serelex.cental.be`
- **Метрики основанные на Веб корпусе:** `http://cwl-projects.cogsci.rpi.edu/msr`
- **LSA:** `http://lsa.colorado.edu`
- **DefVectors:** `http://github.com/jgc128/defvectors`

План

Публикации

- Panchenko A., Morozova O., Naets H. **A Semantic Similarity Measure Based on Lexico-Syntactic Patterns.** In Proceedings of KONVENS 2012, pp.174–178, 2012
- Panchenko A., Romanov P., Morozova O., Naets H., Philippovich A., Fairon C. **Serelex: Search and Visualization of Semantically Related Words**. In Proceedings of the 35th European Conference on Information Retrieva (ECIR 2013).
- Панченко А., Романов П., Романов А., Филиппович А., Филиппович Ю., Морозова О. **Серелекс: поиск и визуализация семантически связанных слов**. (АИСТ 2013)

Демо

- http://serelex.cental.be/

ferrari Search

Results count: 367

1. porsche
2. maserati
3. lamborghini
4. marque
5. bmw
6. audi
7. mercedes
8. jaguar
9. mclaren
10. dodge
11. bugatti
12. lancia
13. lotus
14. isuzu
15. tvr
16. vw
17. nissan
18. rally driving
19. fiat
20. mazda

Show all results...

computational linguistics Search

Results count: 88

1. psycholinguistics
2. machine learning
3. computer science
4. knowledge representation
5. cognitive science
6. artificial intelligence
7. information retrieval
8. neuroinformatics
9. natural language
10. graduate student
11. library science
12. distributed computing
13. research community
14. information processing
15. subfield
16. language acquisition
17. computer vision
18. soas
19. sociolinguistics
20. data mining

Show all results...

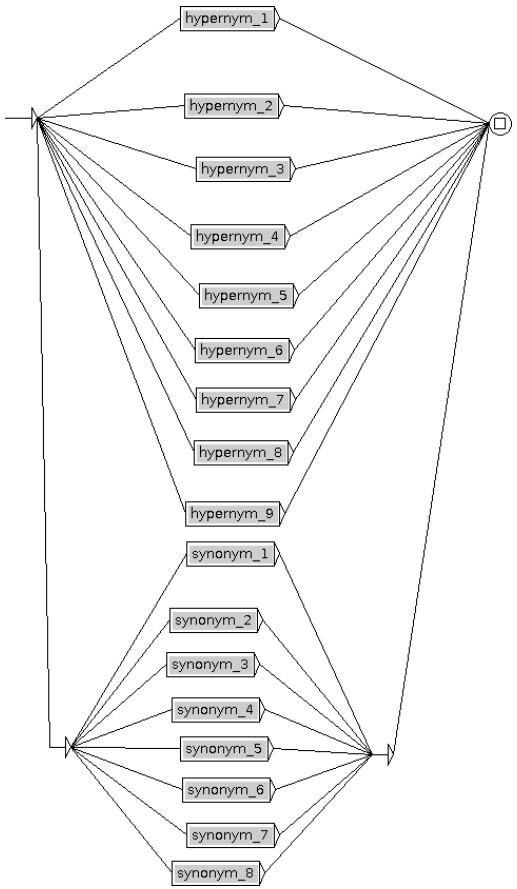
Лексико-синтаксические паттерны

- 18 паттернов извлекающих **гиперонимы**, **ко-гипонимы** и **синонимы**

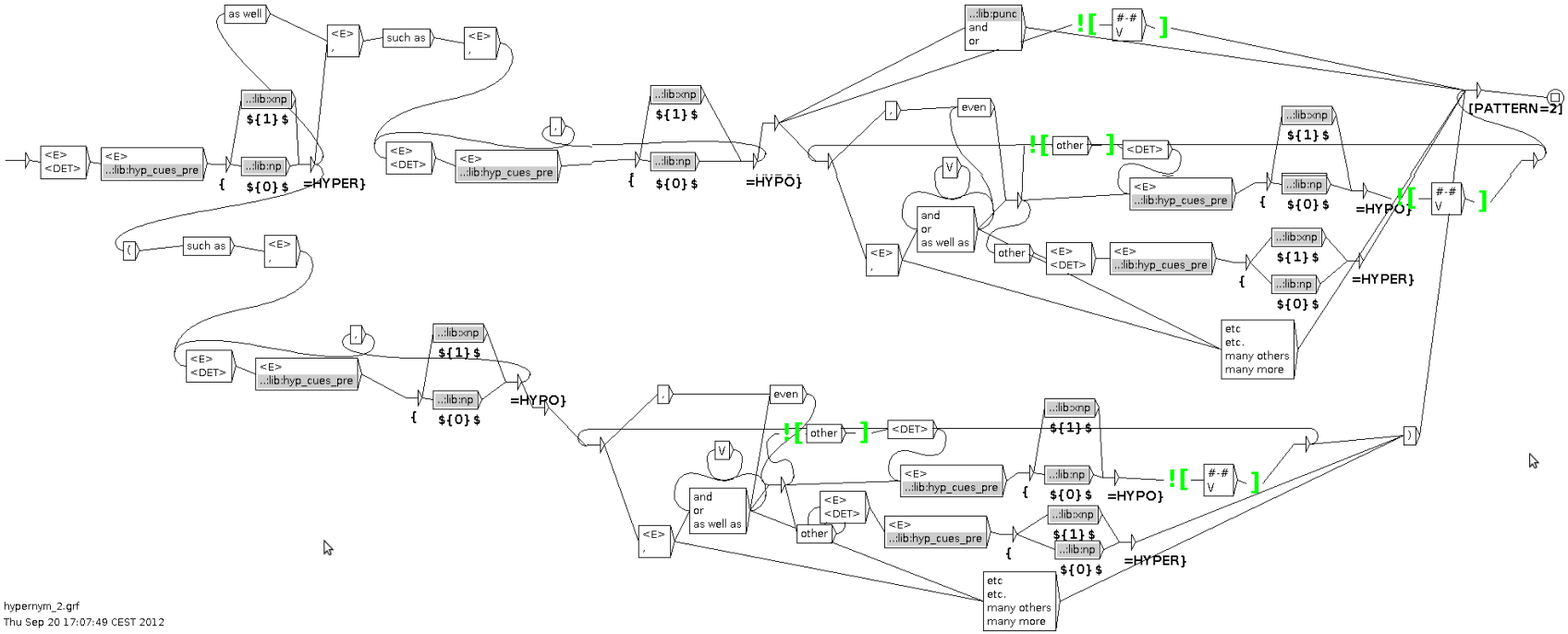
- *such NP as NP, NP[,] and/or NP;*
- *NP such as NP, NP[,] and/or NP;*
- *NP, NP [,] or other NP;*
- *NP, NP [,] and other NP;*
- *NP, including NP, NP [,] and/or NP;*
- *NP, especially NP, NP [,] and/or NP;*
- *NP: NP, [NP,] and/or NP;*
- *NP is DET ADJ.Superl NP;*
- *NP, e. g., NP, NP[,] and/or NP;*
- *NP, for example, NP, NP[,] and/or NP;*
- *NP, i. e.[,] NP;*
- *NP (or NP);*
- *NP means the same as NP;*
- *NP, in other words[,] NP;*
- *NP, also known as NP;*
- *NP, also called NP;*
- *NP alias NP;*
- *NP aka NP.*

Основной каскад автоматов

- Каскад конечных автоматов (FST)
- В формете `Unitex`



Пример реализации паттерна в виде автомата



- Паттерны основанные на автоматах позволяют учесть лингвистическую вариацию сохраняя точность
- В отличие от паттернов основанных на строках (Bollegala et al., 2007)

PatternSim: основные этапы

Корпус Wikipedia+ukWaC: $2.9 \cdot 10^{12}$ токенов

Паттерны извлекают конкордансы

- `such diverse {[occupations]} as {[doctors]}, {[engineers]} and {[scientists]}[PATTERN=1]`
- `such {non-alcoholic [sodas]} as {[root beer]} and {[cream soda]}[PATTERN=1]`
- `{traditional[food]}, such as {[sandwich]},{[burger]}, and {[fry]}[PATTERN=2]`

Количество извлечений

- Wikipedia – 1.196.468
- ukWaC – 2.227.025
- WaCypedia+ukWaC – 3.423.493

Вычисление подобия

Формула Efreq-Rnum-Cfreq-Pnum

$$s_{ij} = \sqrt{p_{ij}} \cdot \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}.$$

- $P(c_i, c_j) = \frac{e_{ij}}{\sum_{ij} e_{ij}}$ – вероятность извлечения отношения $\langle c_i, c_j \rangle$, где e_{ij} – частота взаимной встречаемости слов c_i и c_j во множестве конкордансов
- $P(c_i) = \frac{f_i}{\sum_i f_i}$ – вероятность слова c_i, где f_i – частота c_i
- $b_{i*} = \sum_{j:e_{ij} \geq \beta} 1$ – количество извлечений слова c_i с частотой $\geq \beta$, где $\mu_b = \frac{1}{|C|} \sum_{i=1}^{|C|} b_{i*}$ – среднее количество извлечений для отдельного слова
- $p_{ij} \in [1; 18]$ – количество отдельных паттернов которые извлекли отношение $\langle c_i, c_j \rangle$

Ранжирование семантических отношений

- Точность **сравнима или лучше** чем у аналогов;
- Полнота **меньше** чем у аналогов.

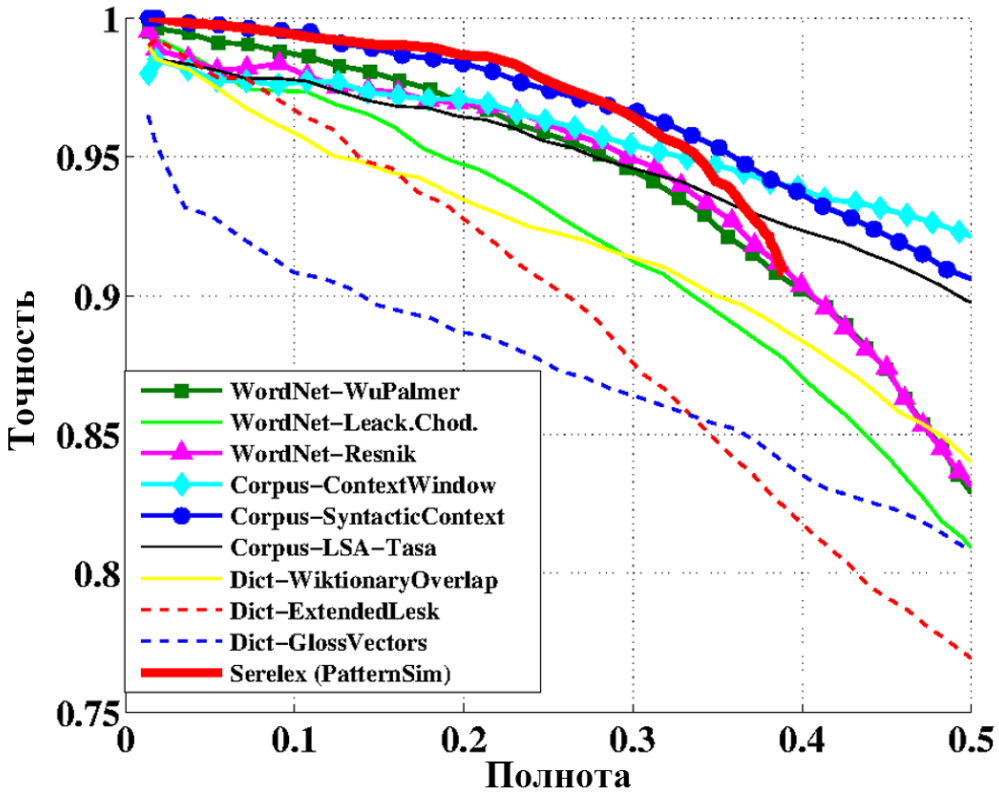
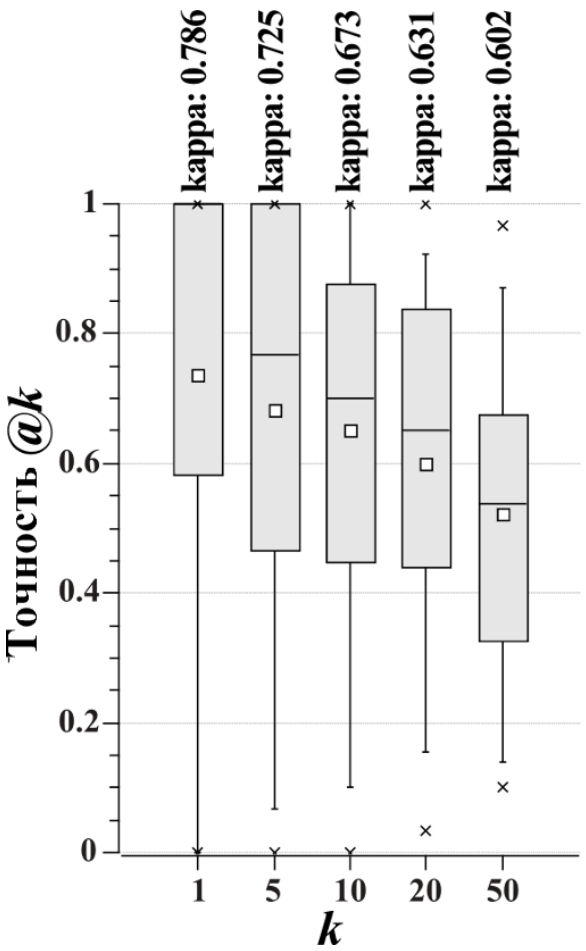


Рис.: График точность-полнота (коллекция BLESS).

Извлечение семантических отношений



- Точность@1 ≈ 0.80;
- “Хорошее” лексическое покрытие:

computational linguistics Search

System finds semantically related words.
For example, cottage cheese

Results count: 88

1 psycholinguistics
2 machine learning
3 computer science
4 knowledge representation
5 cognitive science
6 artificial intelligence
7 information retrieval
8 neuroinformatics
9 natural language
10 graduate student
11 library science

Word to search for: computational linguistics Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show W

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **computational linguistics** (the use of computers applications)
  - *direct hyponym* / *full hyponym*
    - S: (n) machine translation, MT (the use of co one language to another)
  - ***direct hypernym*** / *inherited hypernym* / *sister tern*
    - S: (n) linguistics (the scientific study of langua

План

Публикациии

- Panchenko A., Morozova O. **A Study of Hybrid Similarity Measures for Semantic Relation Extraction.** // Innovative Hybrid Approaches to the Processing of Textual Data Workshop, EACL 2012 — Avignon (France), 2012 — pp. 10–18
- Panchenko A., **Similarity Measures for Semantic Relation Extraction.** PhD thesis. Université catholique de Louvain. 197 pages, 2013, (Chapter 4).
- Panchenko A. **A Study of Heterogeneous Similarity Measures for Semantic Relation Extraction.** // In JEP-TALN-RECITAL 2012 — Grenoble (France), 2012 — pp. 29–42.

Отдельные и гибридные метрики

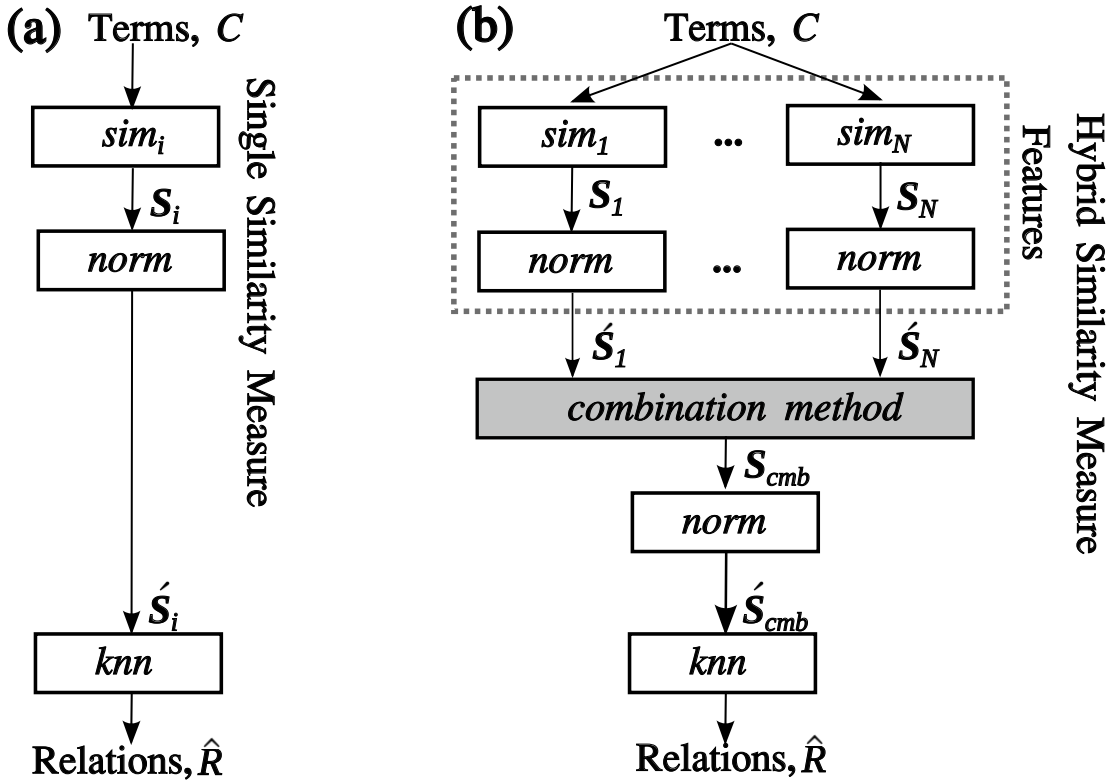


Рис.: Система извлечения семантических отношений основанная на:

- **(a) отдельной** метрике;
- **(b) гибридной** метрике.

16 признаков = 16 отдельных метрик

- 5 метрик основанных на **семантических сетях**:
  1. WuPalmer;
  2. Leacock and Chodorow;
  3. Resnik;
  4. Jiang and Conrath;
  5. Lin.
- 3 метрики основанных на **Веб корпусе** (NGD-Yahoo/Bing/Google);
- 5 метрики основанные на **корпусе текстов**:
  - 2 дистрибутивных (BDA, SDA)
  - 1 лексико-синтаксические шаблоны (PatternSim)
  - 2 другие (LSA, NGD-Factiva)
- 3 метрики основанные на **определениях**
  1. ExtendedLesk;
  2. GlossVectors;
  3. DefVectors-WktWiki.

Методы комбинирования

8 **Logit**, **Logit-L1**, **Logit-L2**.

- Бинарная **логистическая регрессия**;
- **Положительные обучающие примеры** – синонимы, гиперонимы, ко-гипонимы из BLESS/SN;
- **Отрицательные обучающие примеры** – случайные пары семантически несвязных слов из BLESS/SN;
- Отношение $\langle c_i, t, c_j \rangle \in R$ представлено с помощью **вектора попарных близостей**: $\mathbf{x} = (s_{ij}^1, \ldots, s_{ij}^N), N = \overline{2, 16}$;
- Категория y_{ij}:

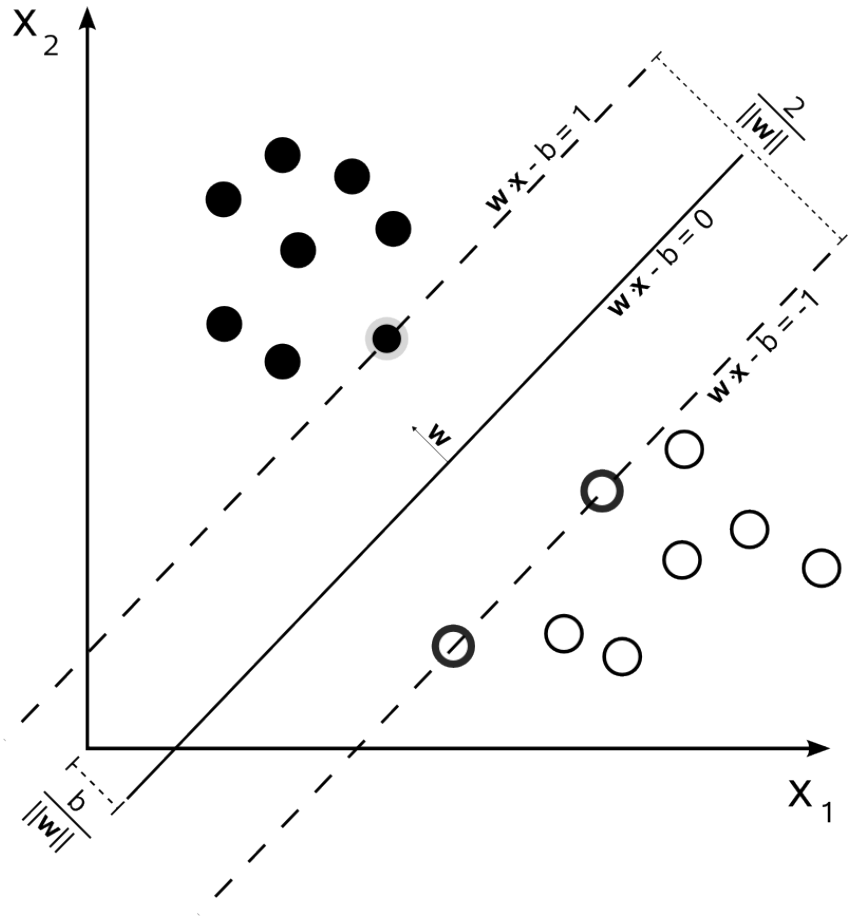
$$y_{ij} = \begin{cases} 0 & \text{если } \langle c_i, t, c_j \rangle \text{ случайное отношение} \\ 1 & \text{иначе} \end{cases}$$

- **Использование модели** $(w_1, \ldots, w_K)$ для комбинирования:

$$s_{ij}^{cmb} = \frac{1}{1 + e^{-z}}, z = \sum_{k=1}^{K} w_k s_{ij}^k + w_0.$$

Методы комбинирования

9 SVM.



- Веса **w** и опорные вектора SV:

$$\mathbf{w} = \sum_{x_i \in SV} \alpha_i y_i \mathbf{x}_i.$$

- **Использование модели**

$$s_{ij}^{cmb} = \mathbf{w}^T \mathbf{x} + b = \sum_{k=1}^{K} w_i s_{ij}^k + b.$$

Методы комбинирования с учителем

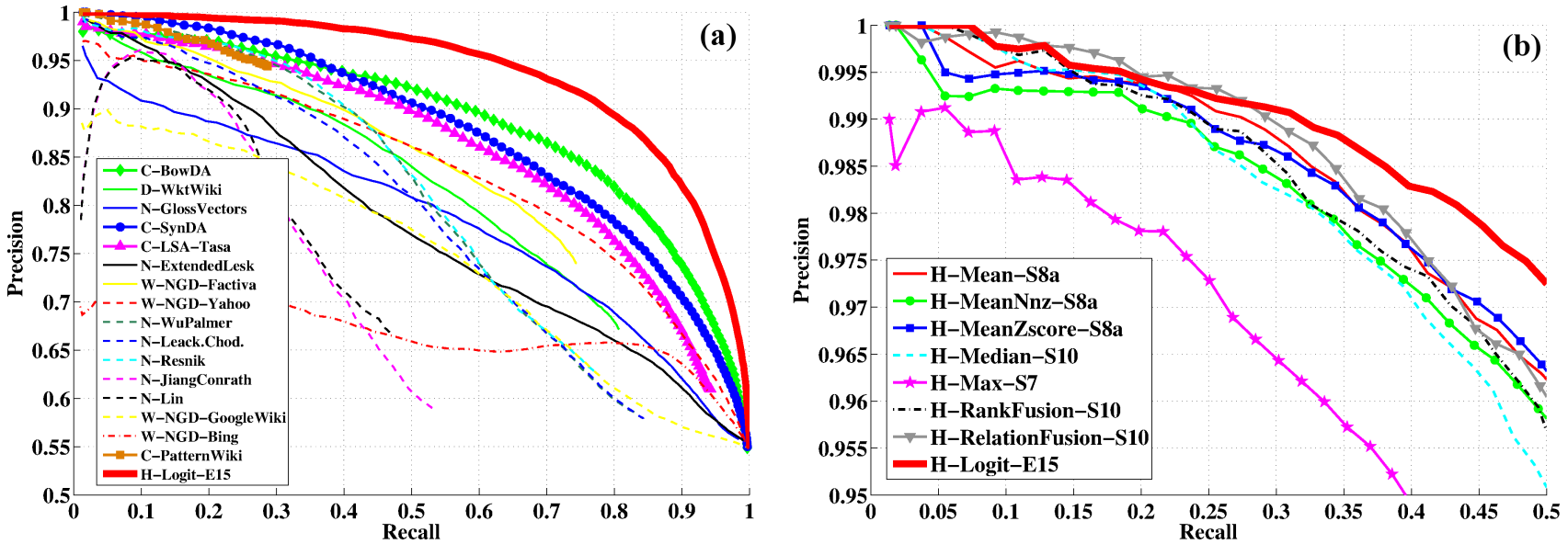


График Точность-Полнота вычисленный на коллекции BLESS:

- **(a)** 16 отдельных метрик и гибридная метрика Logit-E15;
- **(b)** 8 гибридных метрик.

Методы комбинирования с учителем Logit-E15

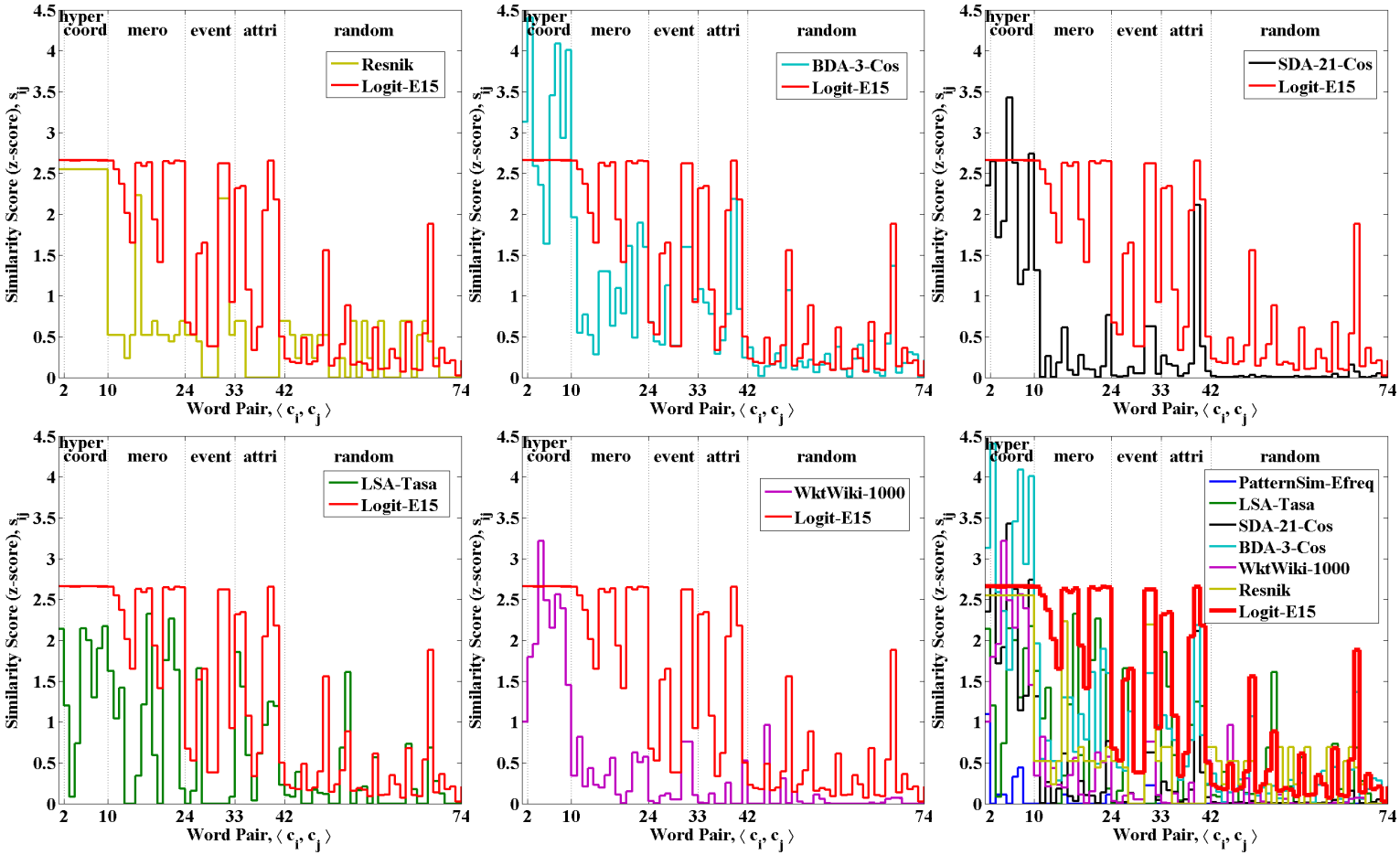


Рис.: Значение подобия между 74 словами связанными со словом “acacia”.

Методы комбинирования с учителем

	BLESS					SN				
Similarity Measure	*Accu.*	*P(10)*	*P (20)*	*P(50)*	*R(50)*	*Accu.*	*P(10)*	*P(20)*	*P(50)*	*R(50)*
C-SVM-linear-E15	**0.833**	0.995	**0.986**	**0.884**	**0.817**	0.820	0.995	0.981	0.816	0.816
C-SVM-poly-E15	0.749	0.993	0.976	0.798	0.737	0.795	0.993	0.977	0.791	0.791
C-SVM-radial-E15	0.832	**0.996**	**0.986**	0.883	0.816	**0.831**	0.995	**0.988**	**0.838**	**0.839**
C-SVM-sigmoid-E15	0.829	0.995	0.985	0.881	0.813	0.811	0.995	0.986	0.807	0.808
ν-SVM-radial-E15	0.827	0.995	0.985	0.879	0.812	0.815	**0.996**	0.984	0.811	0.811
ν-SVM-linear-E15	0.819	**0.996**	0.984	0.877	0.810	0.805	0.994	0.984	0.803	0.803
ν-SVM-poly-E15	0.827	**0.996**	0.985	0.879	0.812	0.826	0.995	**0.988**	0.833	0.833
ν-SVM-sigmoid-E15	0.827	0.995	0.984	0.878	0.811	0.811	0.995	0.984	0.809	0.809
Logit-E15	0.831	0.994	**0.986**	**0.884**	**0.817**	0.823	0.994	0.983	0.819	0.819
LogitL2-E15	0.823	0.995	0.982	0.874	0.808	0.773	0.990	0.967	0.798	0.798
LogitL1-E15	0.824	0.994	0.984	0.874	0.807	0.787	0.992	0.975	0.805	0.805
Logit-E5	0.796	0.989	0.977	0.853	0.788	0.795	0.985	0.965	0.791	0.791
C-SVM-radial-E5	0.802	0.990	0.976	0.857	0.792	0.788	0.980	0.959	0.787	0.787
Logit-E9	0.821	0.991	0.983	0.877	0.810	0.821	0.995	0.982	0.824	0.824
C-SVM-radial-E9	0.824	0.993	0.983	0.875	0.809	0.831	**0.997**	0.988	0.837	0.837
Logit-E15	**0.831**	**0.995**	**0.986**	**0.884**	**0.817**	**0.832**	0.995	**0.989**	**0.840**	**0.839**
C-SVM-radial-E15	**0.831**	0.994	**0.986**	**0.884**	**0.817**	0.823	0.994	0.983	0.819	0.819
C-SVM-radial-E15 ($C = 32, \gamma = 2$)	**0.855**	0.987	0.979	**0.900**	**0.831**	**0.846**	0.983	0.981	**0.846**	**0.846**
C-SVM-radial-E15 ($C = 32, \gamma = .125$)	0.841	**0.996**	**0.987**	0.892	0.824	0.844	**0.995**	**0.990**	0.845	0.845

Методы комбинирования с учителем (продолжение)

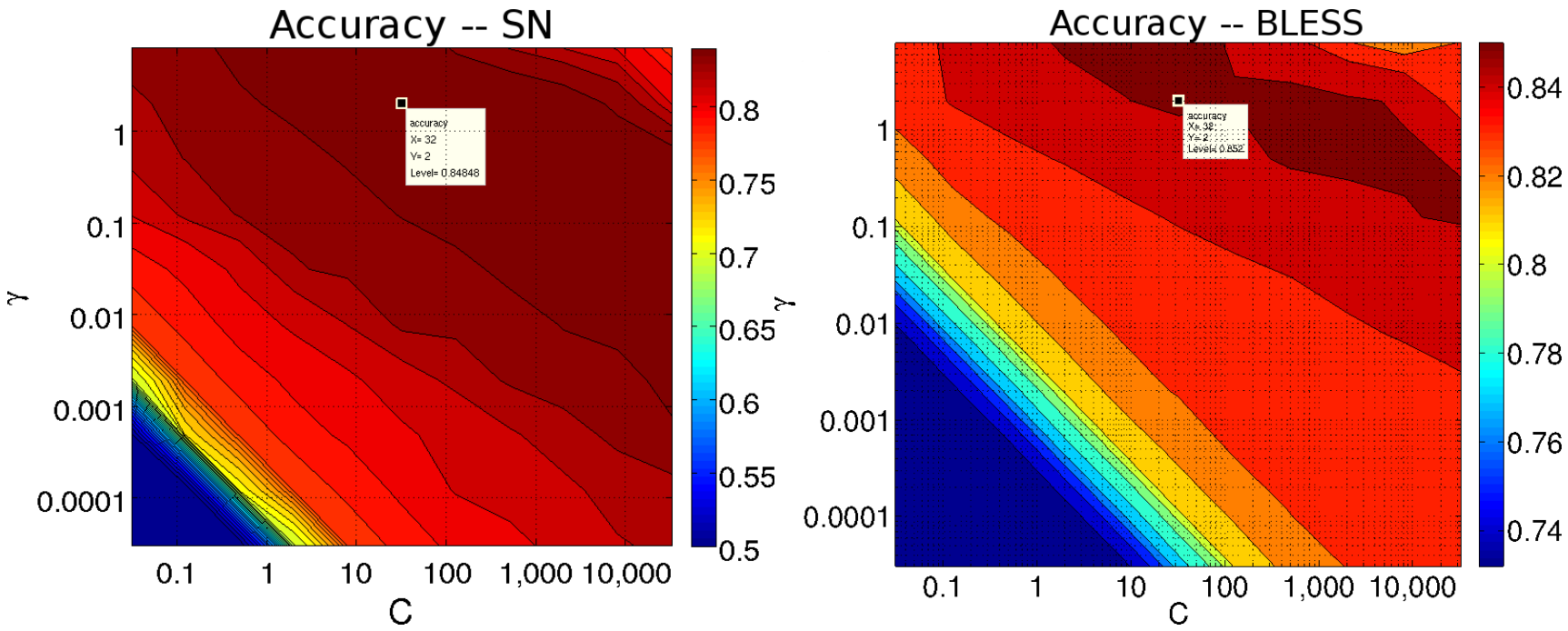


Рис.: Оптимизация мета-параметров метрики C-SVM-radial-E15.

План

План

Серелекс: результаты в виде списка и графа слов

- http://serelex.cental.be/

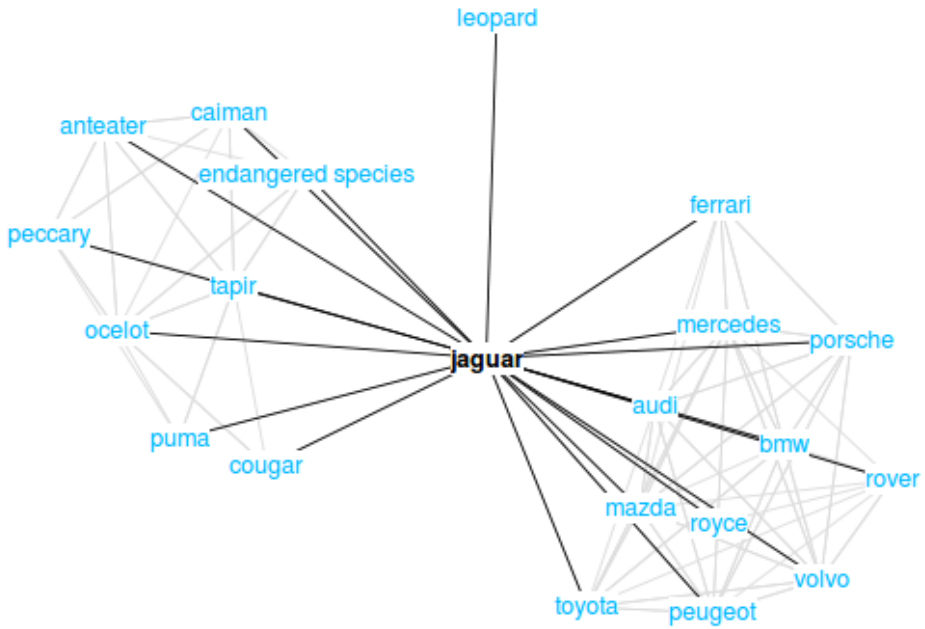
jaguar Искать

Система находит семантически связанные слова.
Например, fedora

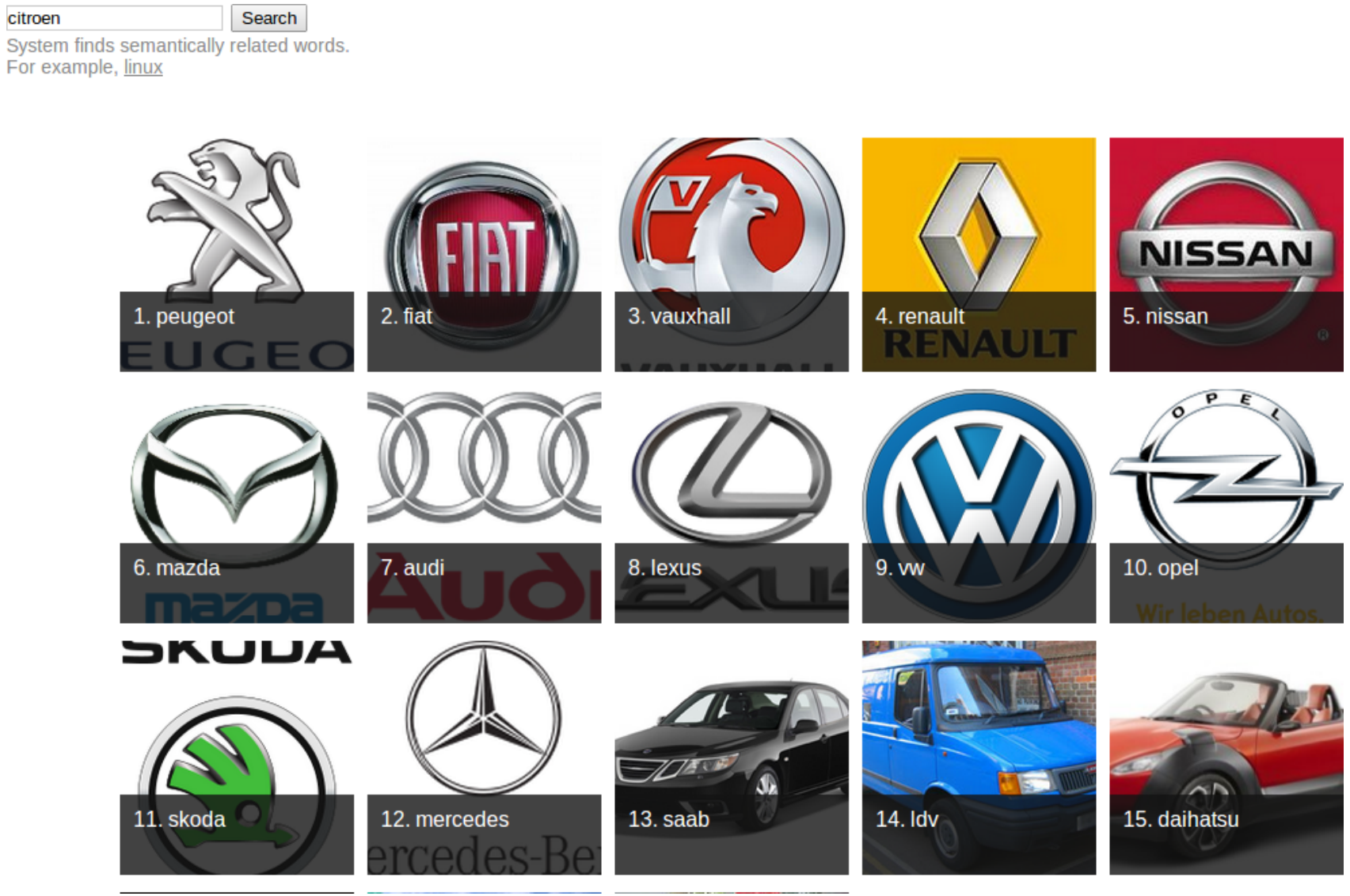
Количество результатов: 685

1. tapir
2. ocelot
3. puma
4. porsche
5. anteater
6. audi
7. cougar
8. mazda
9. rover
10. bmw
11. volvo
12. caiman
13. endangered species
14. ferrari
15. peugeot
16. toyota
17. leopard
18. mercedes
19. peccary
20. royce

Следующие 20 результатов



Серелекс: результаты в виде множества изображений



Оценка качества работы системы Серелекс

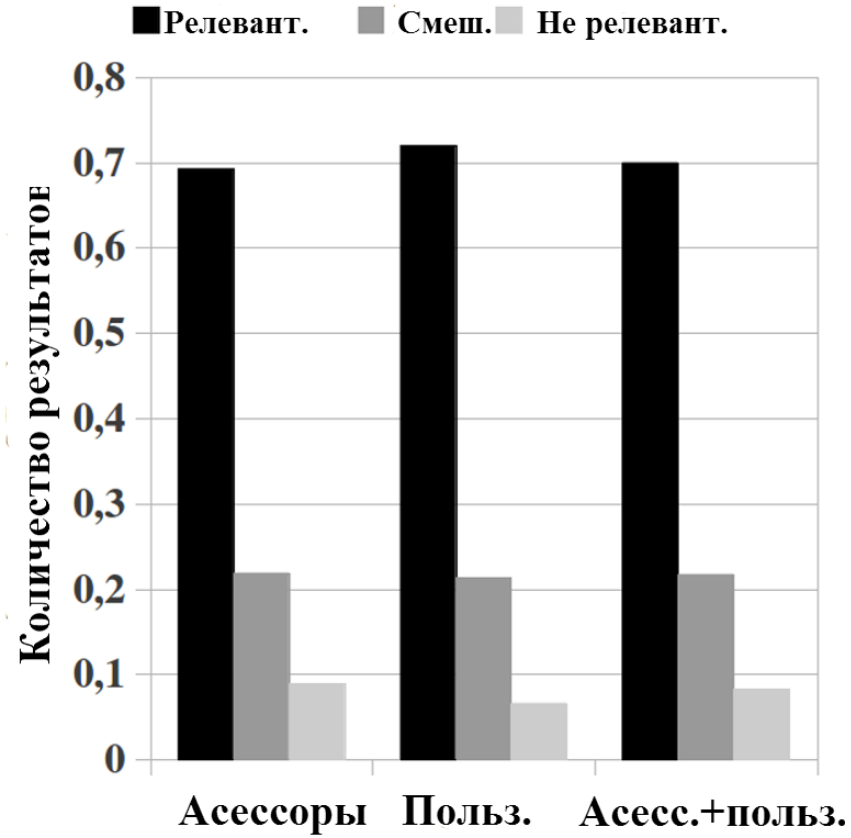


Рис.: Удовлетворенность пользователей первыми 20 результатами поиска для 594 запроса (23 ассесора и 109 пользователей).

План

iCop: классификация имен файлов

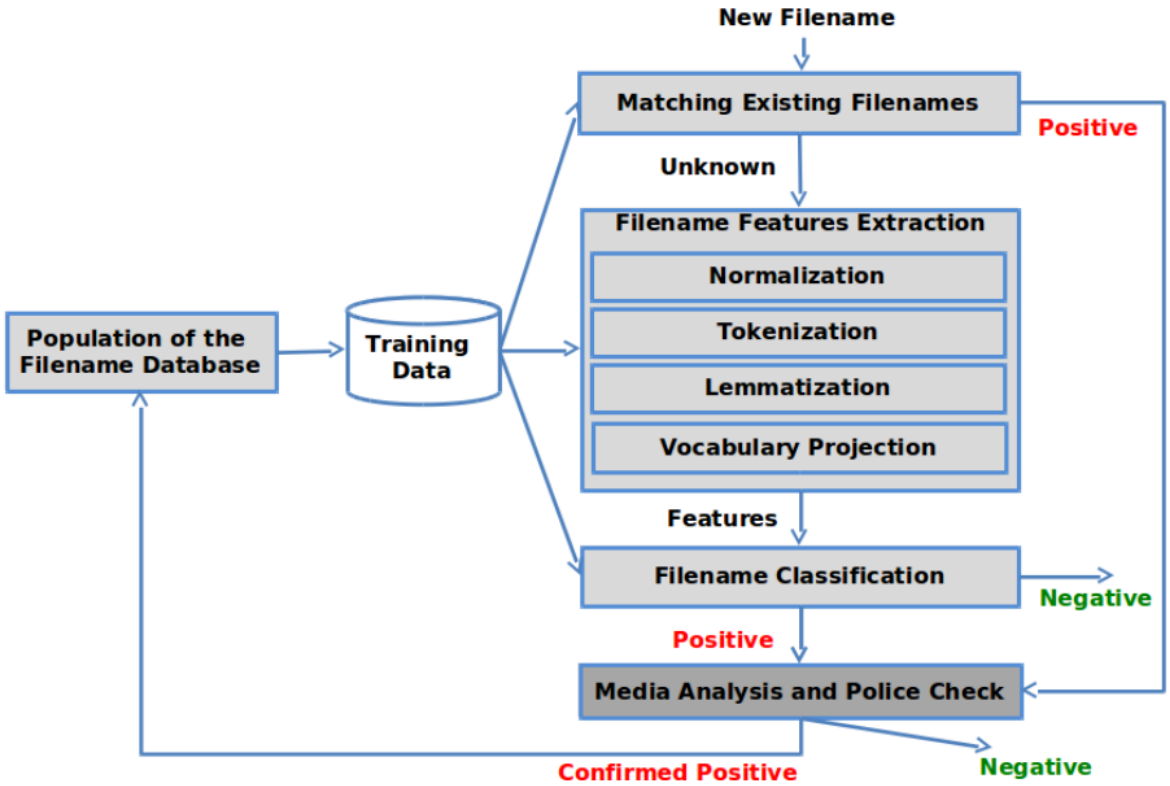
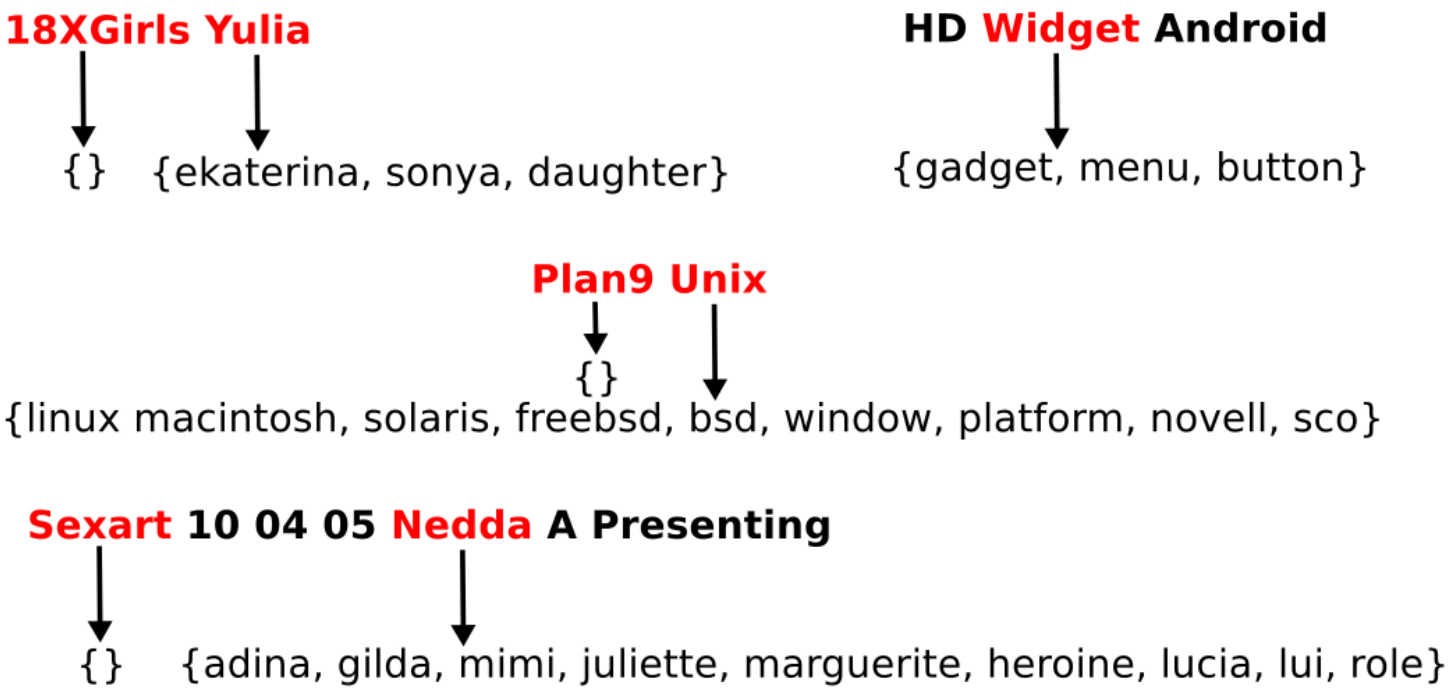


Рис.: Структура системы.

- Использование семантических отношений для расширения имени файла (Vocabulary Projection).

iCop: пример Vocabulary Projection



Качество классификации

Обучающая выборка	Тестовая выборка	Accuracy	Accuracy (voc. projection)
Gallery (train)	Gallery	96.41	**96.83** (+0.42)
PirateBay Title+Desc+Tags	PirateBay Title+Desc+Tags	**98.92**	98.86 (–0.06)
PirateBay Title+Tags	PirateBay Title+Tags	**97.73**	97.63 (–0.10)
Gallery	PirateBay Title+Desc+Tags	90.57	**91.48** (+0.91)
Gallery	PirateBay Title+Tags	84.23	**88.89** (+4.66)
PirateBay Title+Desc+Tags	Gallery	88.83	**89.04** (+0.21)
PirateBay Title+Tags	Gallery	91.16	**91.30** (+0.14)

Таблица: Качество классификации с использованием C-SVM-linear с учетом кросс-валидации.

Спасибо за внимание!
Вопросы?