

Измерение семантической близости концепций Википедии, основанное на анализе ссылок между статьями

Денис Турдаков
ВМК МГУ / ИСП РАН

Московская секция *ACM SIGMOD*, 27 марта 2008 г.

План доклада

- Меры близости в графах
- Граф ссылок Википедии
- Эффективное измерение семантической близости
- Эксперименты и оценки
- Приложения

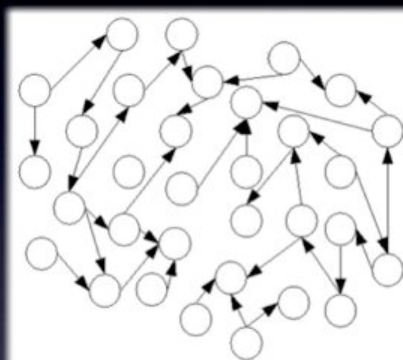
Меры семантической близости в графах

- Базовые меры
 - Мера Дайса
 - Мера Жаккарда
- Теоретико-графовые
 - SimRank

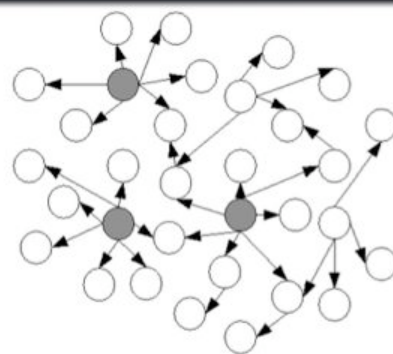
$$Dice(A, B) = \frac{2 \times |n(A) \cap n(B)|}{|n(A)| + |n(B)|}$$

$$sim_{\ell+1}(u, v) = \frac{c}{|I(u)| \cdot |I(v)|} \sum_{u' \in I(u)} \sum_{v' \in I(v)} sim_{\ell}(u', v')$$

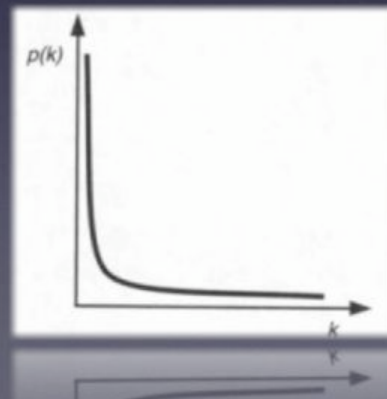
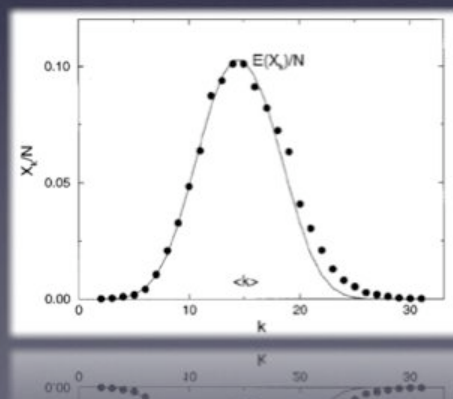
Безмасштабные графы. Статистические свойства



(a) Random network

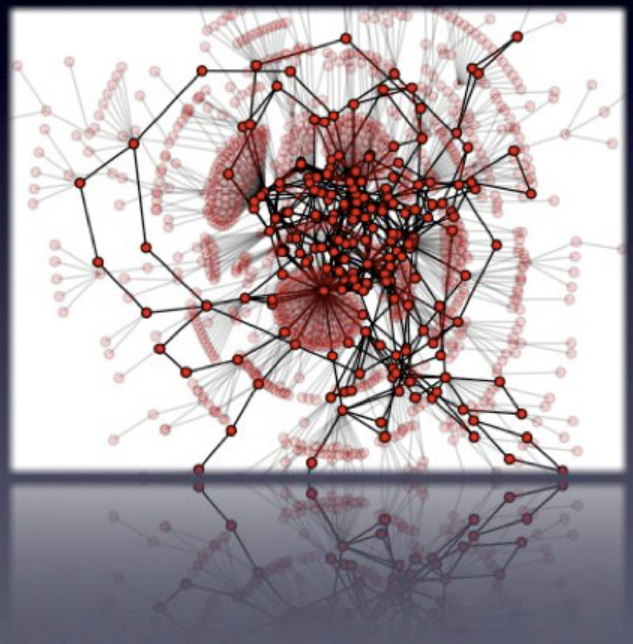


(b) Scale-free network



Безмасштабные графы. Статистические свойства

- Хабы, которые связывают сеть
- Большое ядро
- Малый диаметр - $\ln(N)$
- Само-похожесть, свойства не зависят от размеров сети



Википедия

- 2 304 627 статей
- Все время обновляется
- Граф ссылок - безмасштабный
- Имеет высокое качество

Википедия

Moscow State University

From Wikipedia, the free encyclopedia

Coordinates: 55°42′10″N, 37°31′50″E

"MGU" redirects here. For other uses, see MGU (disambiguation).

M.V. Lomonosov Moscow State University (Russian: Московский государственный университет имени М.В.Ломоносова, romanised: *Moskovskiy gosudarstvennyy universitet imeni M.V. Lomonosova*, often abbreviated **МГУ**, **MSU**, **MGU**) is the largest university in Russia. Founded in 1755, it claims to be the oldest university in Russia. As of 2004, the university has some 4,000 staff teaching 31,000 students and 7,000 postgraduates. Its current rector is Viktor Sadovnichiy.

Contents [hide]

- History
- The campus
- Faculties
- Institutions and research centres
- Famous alumni and faculty
- Partner universities
- See also
- References
- External links

History

[edit]



This article or section's coverage of a controversial issue may be inaccurate or unbalanced in favor of certain viewpoints. Please improve the article by adding information on neglected viewpoints, or remove the tag if the article is about a well-established fact.

M.V. Lomonosov Moscow State University

Московский государственный университет имени М.В.Ломоносова



Moscow State University's main building

Established:	25 January 1755
Rector:	Viktor Sadovnichiy
Faculty:	4,000
Staff:	15,000
Undergraduates:	40,000
Postgraduates:	7,000
Location:	Moscow, Russia

Location: Moscow, Russia
Postgraduates: 7,000

Википедия

Main article: List of Moscow State University people

Moscow State University has produced a number of Nobel laureates, Fields Medal winners and Heads of states.

Partner universities

[edit]

-  Ching Yun University, Taiwan

See also

[edit]

- Seven Sisters (Moscow)
- Education in Russia
- List of universities in Russia

References

[edit]

- *History of Moscow University* – from the official website of MSU.

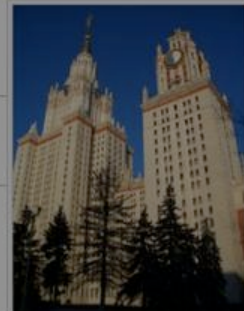
External links

[edit]

- *Moscow State University*
- *Moscow State University* campus on Google Maps
- (Russian) *MechMath* faculty - Unofficial website
- (Russian) *CMC* faculty - Unofficial website
- (Russian) *Journalism* faculty - Unofficial website
- (Russian) 1950*ies* film about constructing of the University



Wikimedia Commons has media related to:
Moscow State University



MSU main building



v · d · e

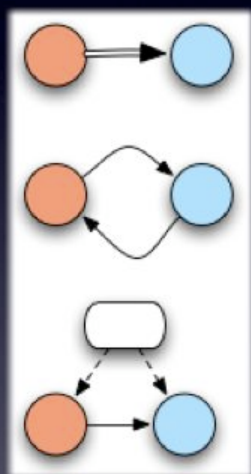
Institutional Network of the Universities from the Capitals of Europe

[hide]

Tirana (TU) • Tirana (PUT) • Wien • Brussels (VUB) • Brussels (ULB) • Sofia • Zagreb • Nicosia • Prague • Copenhagen • Tallinn (TU) • Tallinn (TUT) • Helsinki • Paris I • Paris III • Paris VI • Paris-Dauphine • Berlin (FUB) • Berlin (HUB) • Athens • Budapest • Dublin • Rome-La Sapienza • Rome-Tor Vergata • Rome-Tre • Riga • Vilnius • Skopje • Amsterdam • Oslo • Warsaw • Lisboa • Bucharest • **Moscow** • Bratislava • Ljubljana • Madrid (UAM) • Madrid (UCM) • Stockholm • Lausanne • London

Categories: Moscow State University | Stalinist architecture | Visitor attractions in Moscow | 1755 establishments | Educational institutions established in the 1750s | 1780s architecture | 1950s architecture | Education in the Soviet Union | Education in Russia | Skyscrapers between 200 and 249 meters | Buildings and structures built in the Soviet Union | Buildings and structures in Moscow | Lists of universities and colleges | Universities and colleges in Russia | Science and technology in Russia | Russian culture

Типы ссылок



See Also	5	Double Link:	2
Inverse See Also	2	Common Category	1.5
Regular Link	1	Inverse Regular Link	0.5
Date	0.1	Template	0.1

Date	0.1	Template	0.1
------	-----	----------	-----

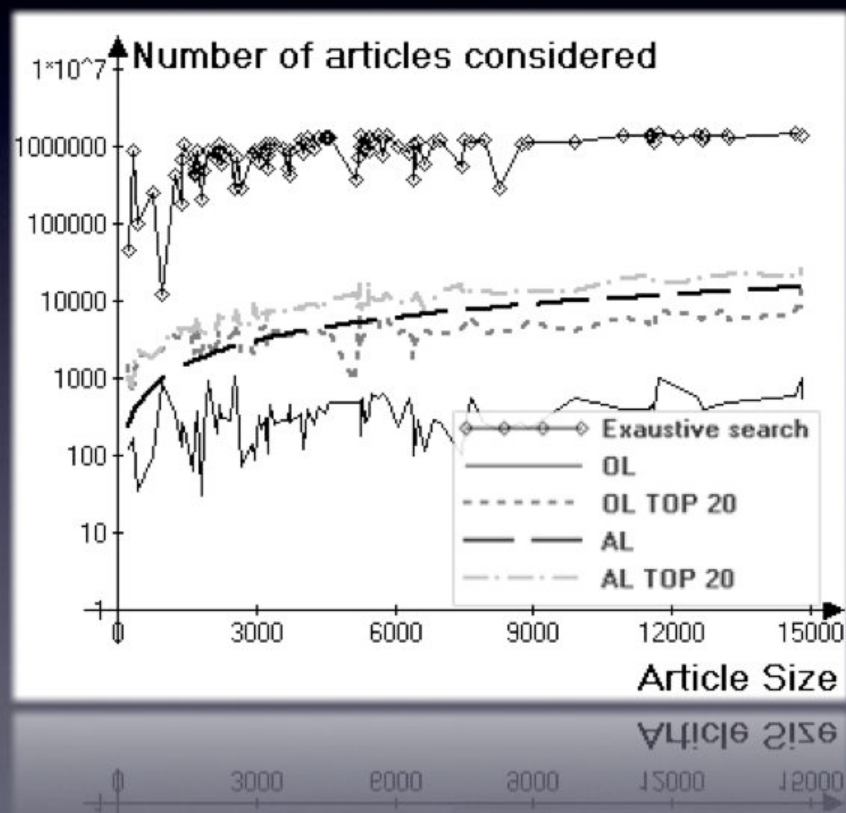
Вычисление близости

- Задача: получить N наиболее близких по смыслу статей
- Статья “United Kingdom” содержит 80000 входящих и исходящих ссылок
- Более 1 000 000 статей с ненулевым значением близости
- Подход “в лоб” не применим

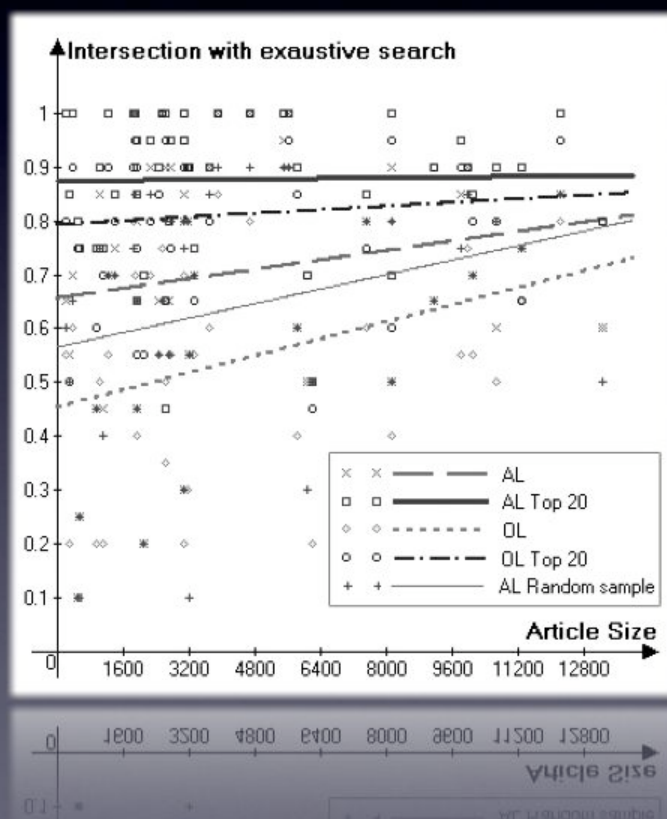
Эвристики

- OL (Outgoing Links) - эффективна для больших статей
- “United Kingdom” содержит 900 исходящих ссылок
- OL Top 20 - основана на наблюдении, что похожие концепции, не имеющие общей ссылки, будут иметь сильные связи с общими соседями
- AL (All Links) и AL Top 20 - для небольших статей

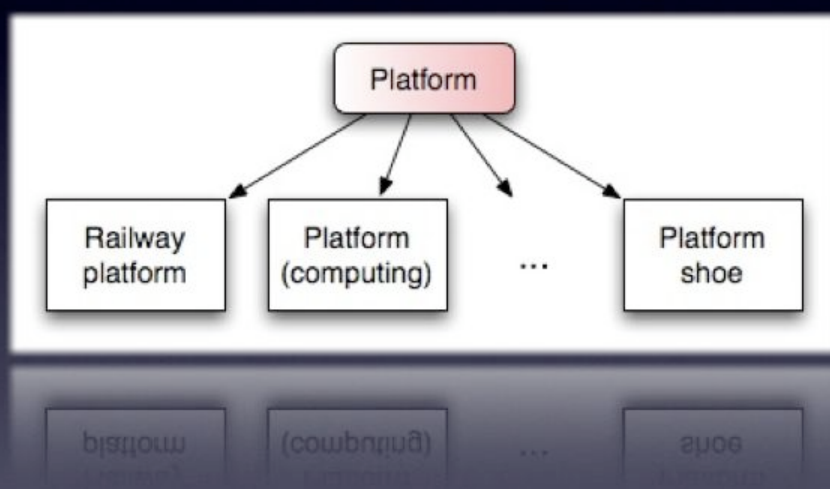
Эвристики. Эффективность



Эвристики. Точность



Устранение семантической неоднозначности

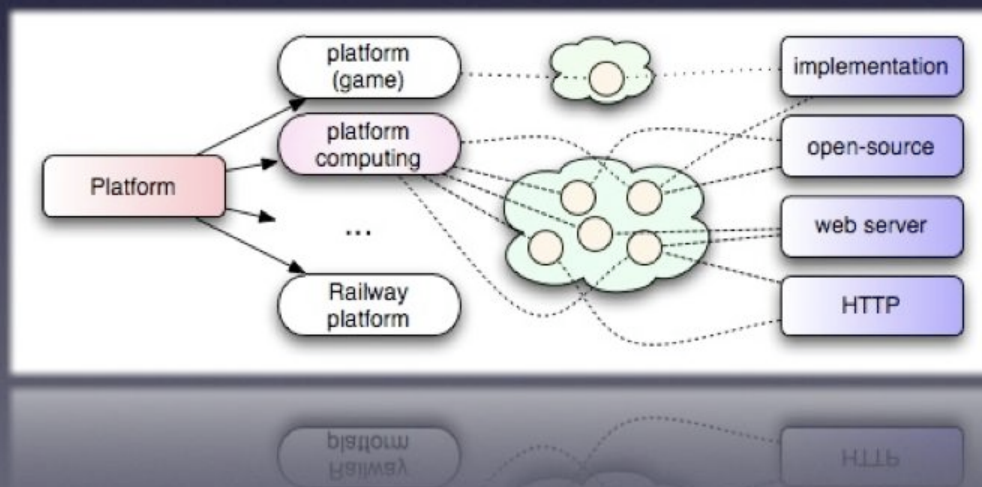


Устранение семантической НЕОДНОЗНАЧНОСТИ

- Википедия - источник многозначных терминов
 - Страницы устранения неоднозначности:
 - Platform (disambiguation) содержит 16 значений термина
 - Каждый вариант представлен ссылкой на статью с его описанием
- В википедии более 80 000 таких статей
- Страницы и редиректы с одинаковыми названиями в верхнем регистре
 - “SUX” - “Sioux Gateway Airport”
 - “Sux” - “Suxamethonium chloride”

Устранение семантической неоднозначности

- Jigsaw is W3C's open-source project that started in May 1996. It is a web server platform that provides a sample HTTP 1.1 implementation and ...



Улучшенный алгоритм

- Для каждого значения с термина и терминов контекста $t_1 \dots t_n$ посчитать ранжированный список близких по смыслу статей ($R(c)$ и $R(t_i)$ соответственно), используя метрику AL
- Посчитать расстояние между значением термина и контекста:
 - Посчитать объединение $R(T) = U(R(t_i))$, суммируя веса повторяющихся терминов
 - Посчитать значение близости между $R(c)$ и $R(T)$, используя один из методов: сумму произведений, косинус, меру Дайса или Джаккарда

Тестовое множество

- Не существует общепризнанного тестового набора
- В общем случае необходимо использовать лингвистические инструменты
- Идея: использовать структуру ссылок Википедии для создания тестового множества
 - [platformism | platform]
- Для экспериментов мы использовали корпус, содержащий 1000 многозначных терминов

Результаты

	Naive	Semantic Relatedness
Intersection	61,83	71,8
Cosine	59,82	72,52
Dice	61,42	72,58
Jaccard	60,28	72,58

	Top-2	Top-3
Intersection	85,12	89,03
Cosine	87,46	90,86
Dice	86.95	91.38
Jaccard	86.95	91.38

Уточнение запросов

Blog Search

Query

Wall Street

Stock Market Crash

Monetary Policy

Business

Marketing

Bank

Silicon Valley

New York

Glinden

...purpose technology that knows about arrays, you can clobber a system in which tables are used to simulate arrays. [For] Wall Street it's basically a latency arms race. If your infrastructure was built with one-second latency, it's ...

Glinden

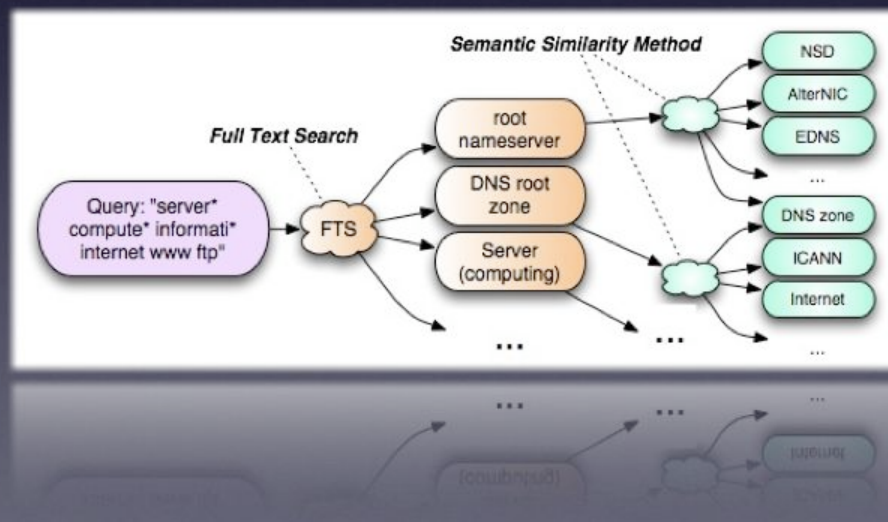
...had us all distracted from the fact that Yahoo is going to report yet another bad quarter. The company told Wall Street analysts that the June quarter is going to come in towards the lower half of revenue and EBITDA ...

Pmarca

...product managers, product marketers, sales, corporate marketing, finance, HR, legal, channels, business development, the strategy team, the international divisions, investors, Wall Street analysts, industry analysts, good press, bad press, press articles being written that you don't know about, customers, ...

Описание области

- Запрос 6 слов
- В результате 11415 терминов, описывающих область



Вопросы?

