

High-Level Visualization of Multilingual Text

Визуализация многоязычного текста на высоком уровне

U.S.-Russia Collaboration for Neurotechnology and
Cognitive Science and Technology

May 10, 2007
10 Мая 2007 г.

Peter Chew Питер Чью
Sandia National Laboratories



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.





Идея

Представьте себе, что вы можете сотворить репрезентацию всего человеческого знания и истории, пользуясь графический или пространственный формат.

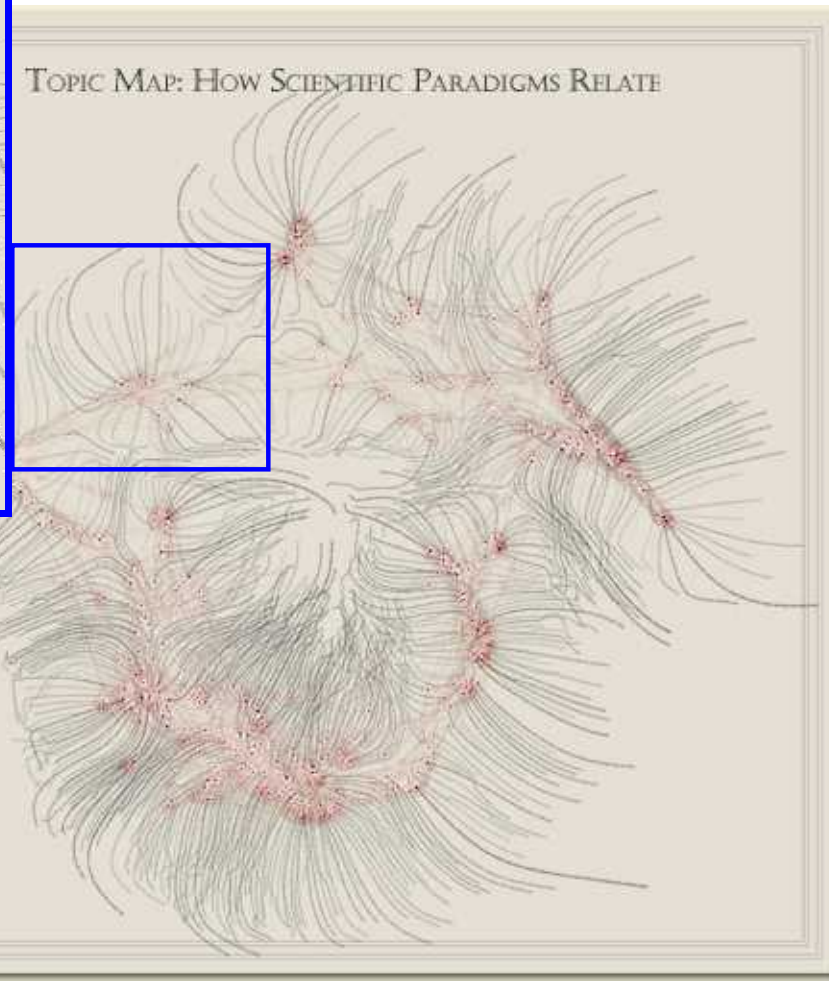
- Как она будет выглядеть?
- Что она будет включать в себе?
- Как она будет употребляться?
- Что люди смогут узнать от нее?
- Будет ли она способна к прогностическому моделированию?
- Как ее построить?



Vision

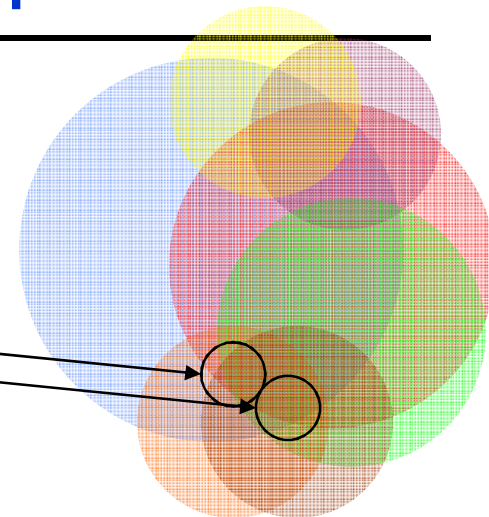
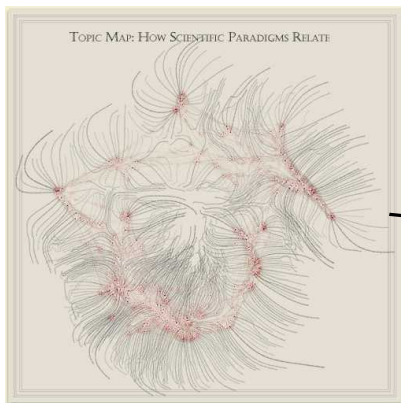
Imagine that you can create a representation, using a graphical or spatial layout, of all of mankind's knowledge and its history.

- What will it look like?
 - What will it contain?
 - How will it be used?
 - What will people be able to learn?
 - Can it be predictive?
-
- How do we make it?



Recent work at Sandia

Недавняя работа в Сандии



ISI database

One year – 1M papers

Relatively clean, fielded data

Matching refs to previous papers

Million node layout

- 100k clusters, very precise topic focus

База данных ISI

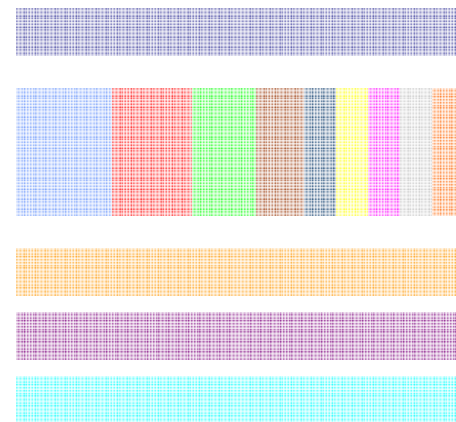
Один год– 1 млн. научных документов

«Чистые» данные, организованные в полях

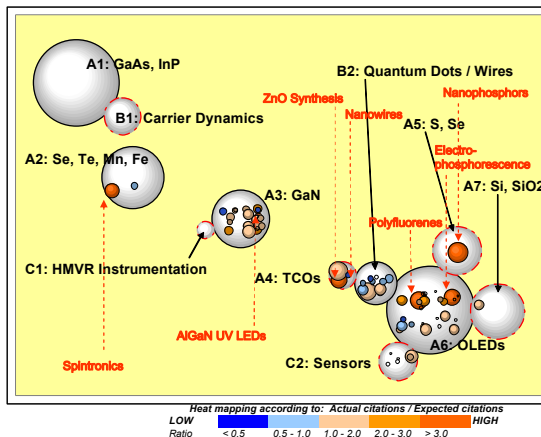
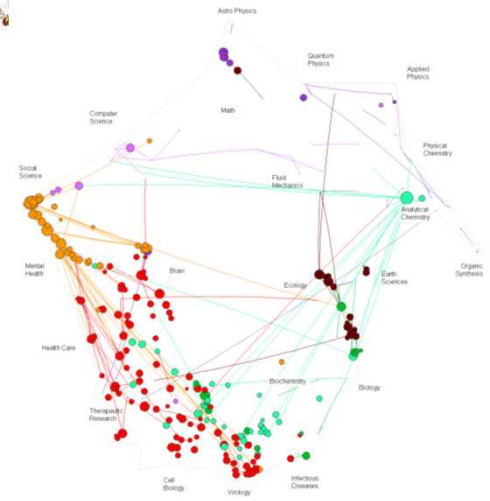
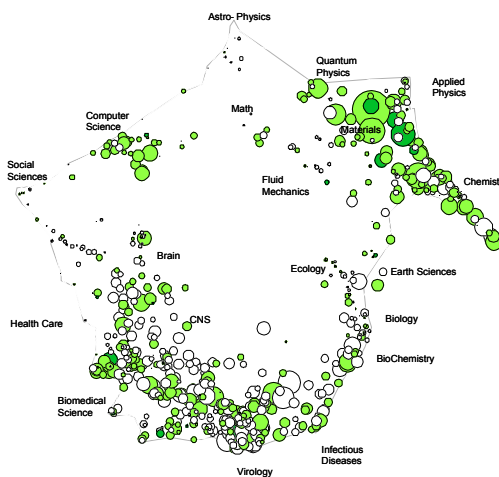
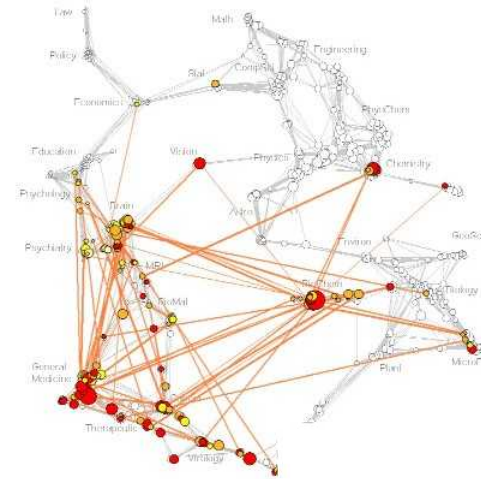
Сходства на основании общих ссылок

График с 1 млн. узлов

- 100 тыс. блоков данных, ясно означающие различные темы



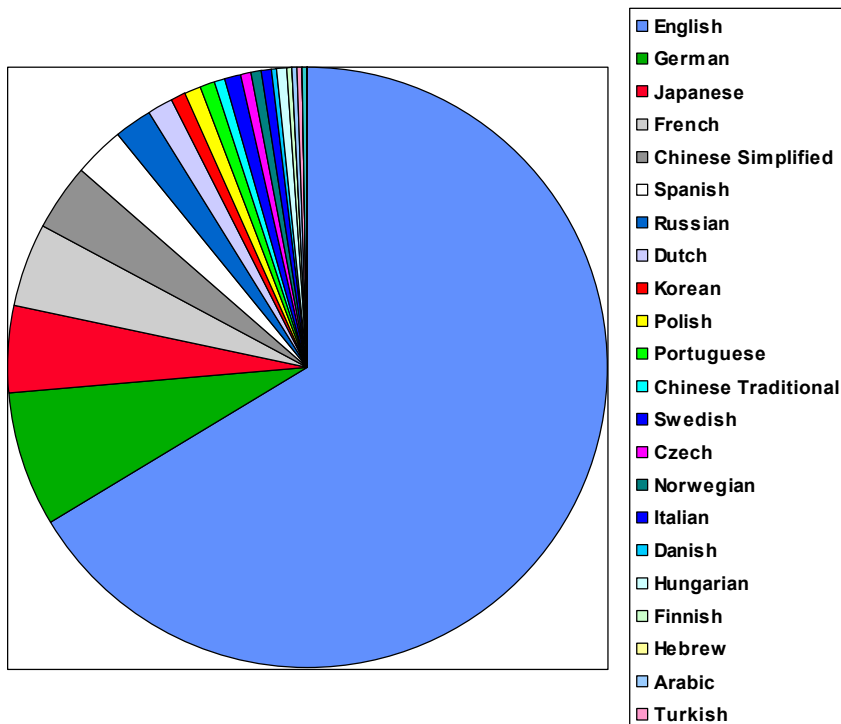
- **Agency, corporate profiles**
- **National strength networks**
- **Collaboration**



Multilingual information retrieval

(identification of similar documents in different languages)

Languages of the internet

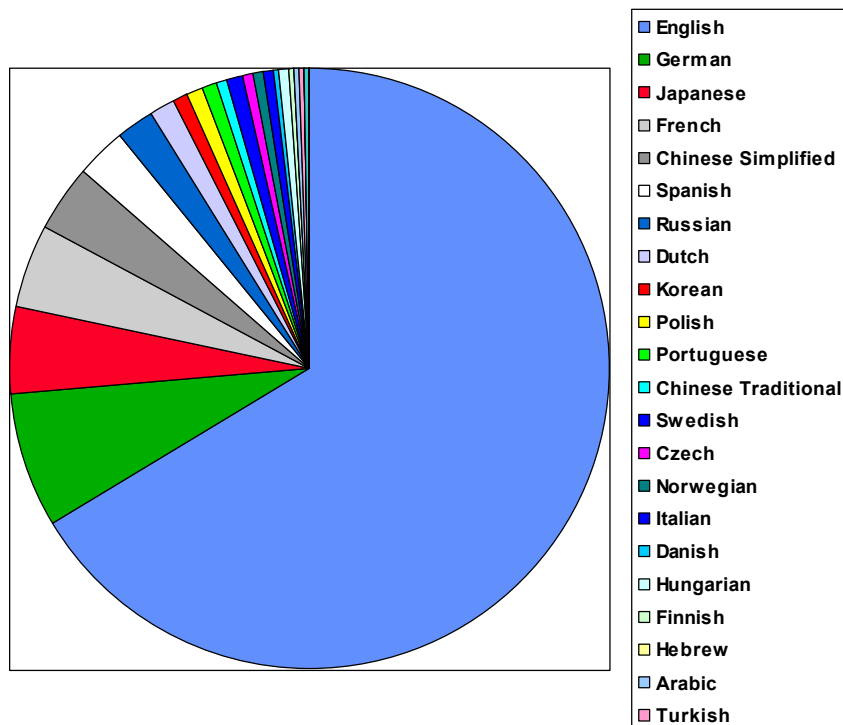


Our framework:

- Identifies similar documents across language boundaries
- Can handle 31 languages (including Japanese)
- Covers about 99.5% of internet content (based on chart at left)
- Is easily extensible to new languages
- Identifies translations of documents with up to 97% precision
- Enables users to find documents of interest in an unknown language

Многоязычный информационный поиск (определение аналогичных документов на разных языках)

Распределение языков в интернете

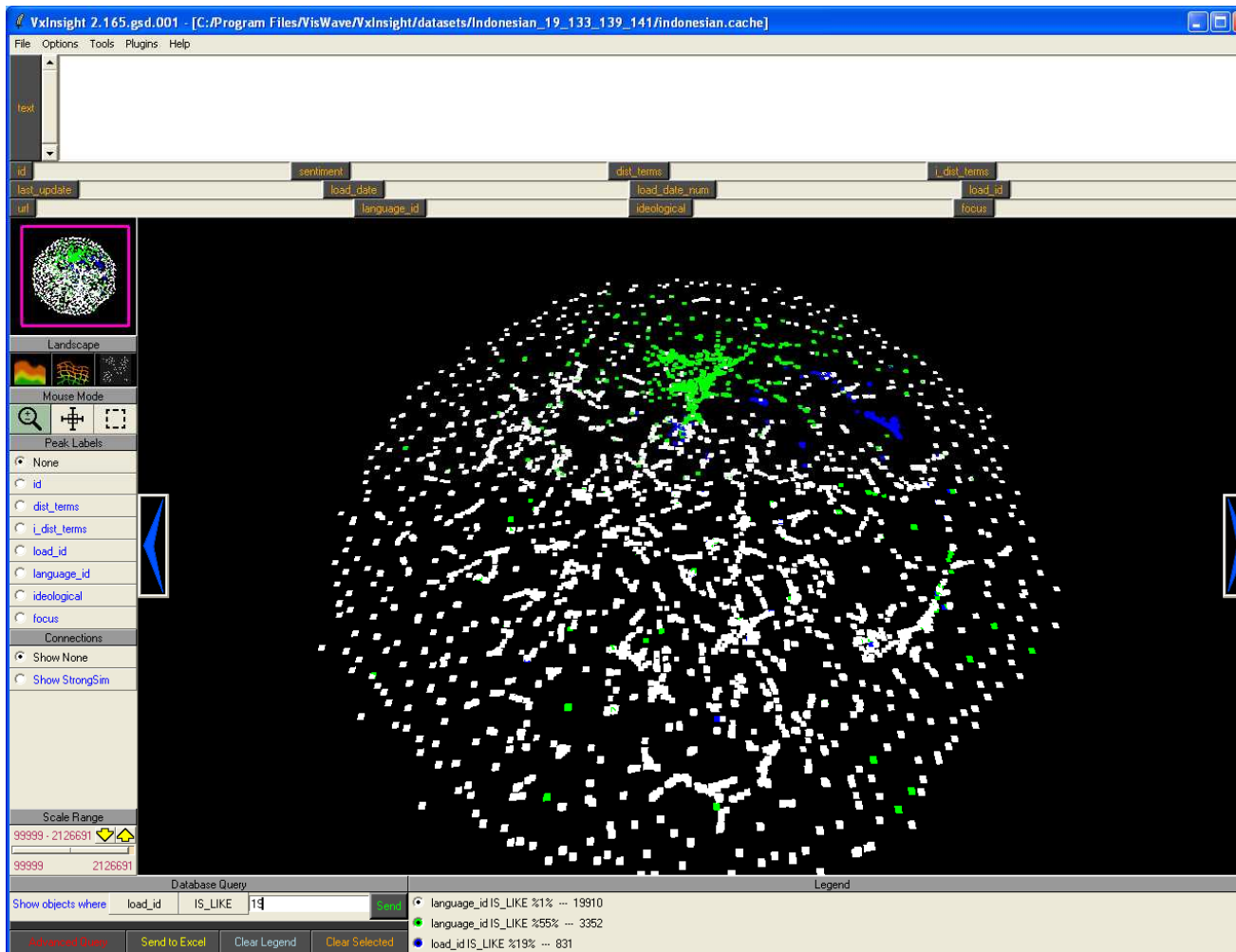


Наша система:

- Определяет аналогичные документы не обращая внимание на их языки
- Может обрабатывать 81 языков
- Охватывает около 99,75% содержания интернета
- Легко растяжимая на другие языки
- Определяет переводы документов с точностью до 99%
- Позволяет пользователям находить документы, представляющий интерес на неизвестном языке

Multilingual visualization

Визуализация на многих языках



Библия как «Розеттский камень»

The Bible as a 'Rosetta Stone'

Язык Language	Текст Text	Слова Words	%
EN	In the beginning God created the heavens and the earth.	10	23.8
RU	В начале сотворил Бог небо и землю.	7	16.7
ES	EN el principio crió Dios los cielos y la tierra.	10	23.8
AR	في البدء خلق الله السموات والارض.	6	14.3
FR	Au commencement Dieu créa les cieux et la terre.	9	21.4
ИТОГО \ TOTAL		42	100.0
Параллельные мини-документы (стихи) Parallel documents (verses)		31226	
Языки с переводом Languages with translations		2426	
Параллельные переводы в электронном формате Electronically-available parallel translations		80+	
Охват содержания интернета Coverage of internet content		99.75%	

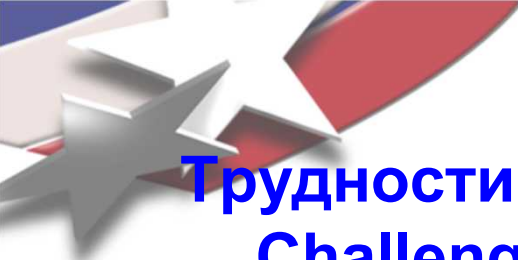


Documents as vectors Документы как векторы

	document/документ			
	1	2	3	
the	2	3	2	7
man	1	1	0	2
walked	1	0	0	1
dog	1	1	1	3
took	0	1	0	1
to	0	1	1	2
park	0	1	1	2
went	0	0	1	1
TOTAL	5	8	6	19

Data can be manipulated using linear algebra, e.g. SVD

Обращение данных линейной алгеброй, нпр. SVD



Трудности с сравнением русского и английского

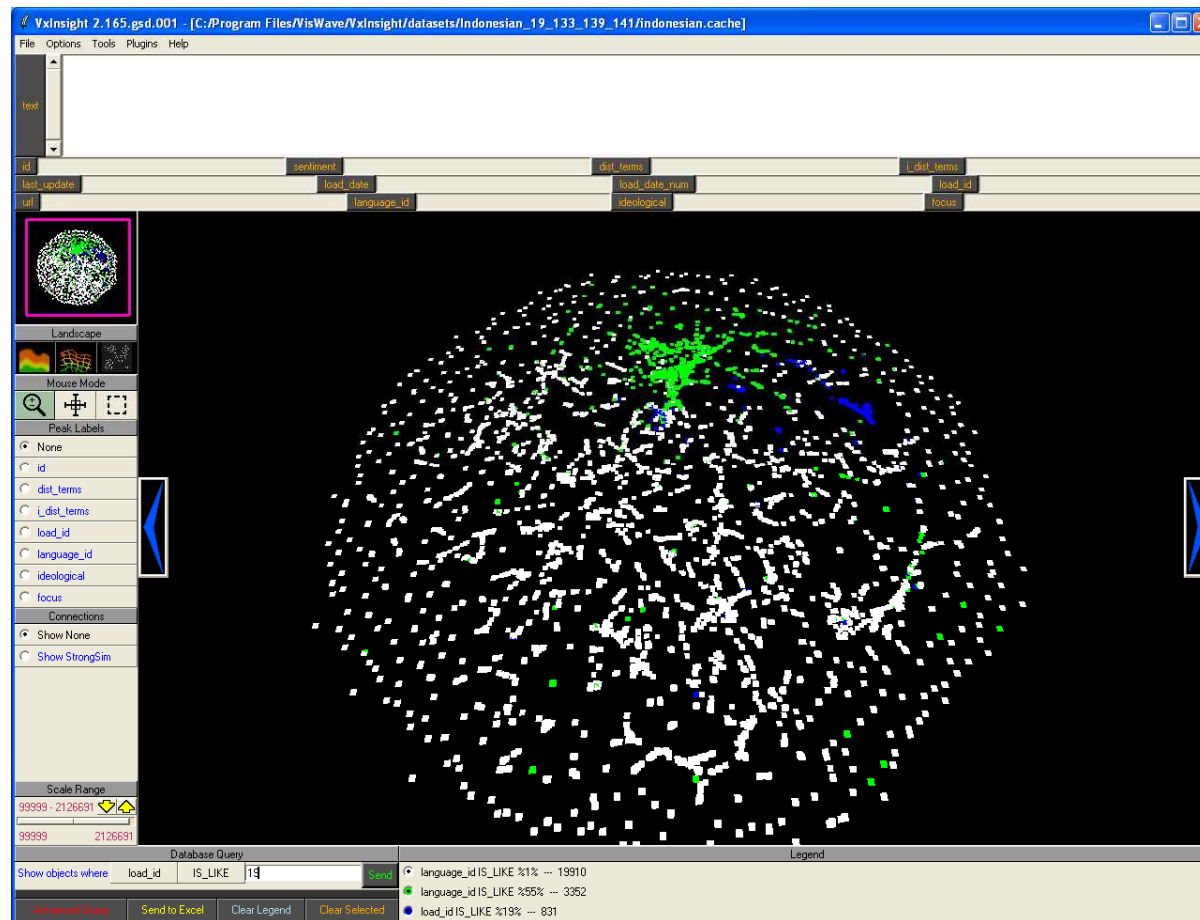
Challenges of comparing Russian and English

Язык / Language	Типы / Types	Экземпляры / Tokens
английский / English	12,335	789,744
русский / Russian	47,226	560,524
испанский / Spanish	28,456	704,004
арабский / Arabic	55,300	440,435
французский / French	20,428	812,947

Analytic and **synthetic** languages
Аналитические и **синтетические** языки

A future project? Будущий проект?

Matching Russian articles to English-language journals
Соответствие документов на русском с журналами на английском





Ссылки и координаты References and contact information

- Chew, Verzi, Bauer and McClain. 2006. Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, 2006, 68–74.
- Chew and Abdelali. 2007, forthcoming. Benefits of the ‘massively parallel Rosetta Stone’: cross-language information retrieval with over 30 languages, *Proceedings of the Association for Computational Linguistics conference*, 2007.
- Chew, Bader, Kolda and Abdelali. Forthcoming. Cross-language information retrieval using PARAFAC2. Submitted to KDD 2007.
- Chew and Abdelali. Forthcoming. The Effects of Language Relatedness and Morphology on Multilingual Information Retrieval: A Case Study With Semitic Languages

SANDIA POINT OF CONTACT:
Peter Chew (pchew@sandia.gov)



DISCUSSION and QUESTIONS

SANDIA POINT OF CONTACT:
Peter Chew (pchew@sandia.gov)