Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

RuSSIR

Russian Summer School
in Information Retrieval

Social Context as Machine Processable Knowledge
Knowledge Based Text Processing

© 2010 Alexander Troussov

The social context (the set of facts or circumstances that surround a situation or event) is hard to represent to our formal reason

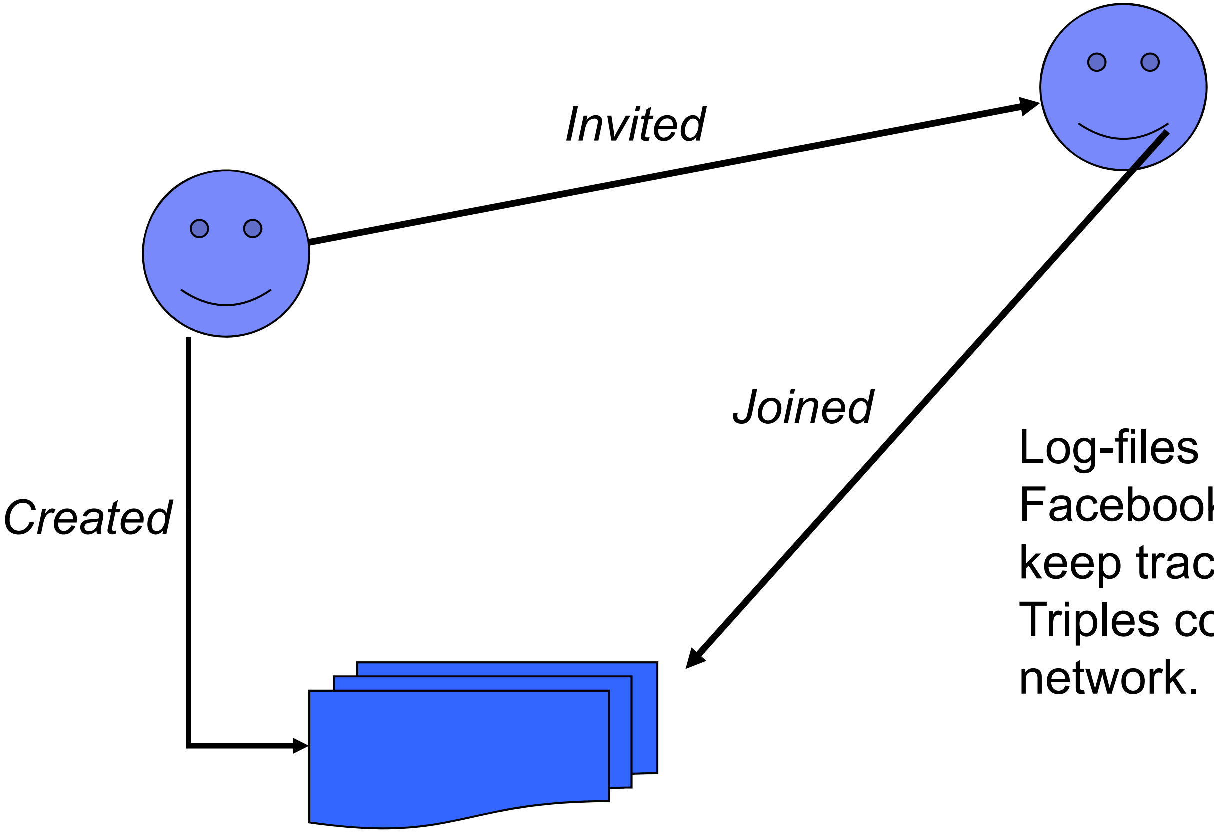
Source: Prof. Noshir Contractor

The social context: introduction

- We live in an increasingly interconnected world of techno-social systems, in which technological infrastructures composed of many layers are interoperating within a social context that drives their everyday use and development. Nowadays, most of the digital content is generated within public systems like Facebook, Delicious, Twitter, blog and wiki systems, and also enterprise environments such as Microsoft SharePoint, and IBM Lotus Connections. These applications have transformed the Web from a mere document collection into a highly interconnected social space where documents are actively exchanged, filtered, organized, discussed and edited collaboratively.
- The emergence of the Social Web opens up unforeseen opportunities for observing social behavior by tracing social interaction on the Web. In these socio-technological systems “everything is deeply intertwingled” using the term coined by the pioneer of the information technologies Ted Nelson[ ]: people are connected to other people and to “non-human agents” such as documents, datasets, analytic tools, tags and concepts. These networks become increasingly multidimensional providing rich context for network mining and understanding the role of particular nodes representing people and digital content.

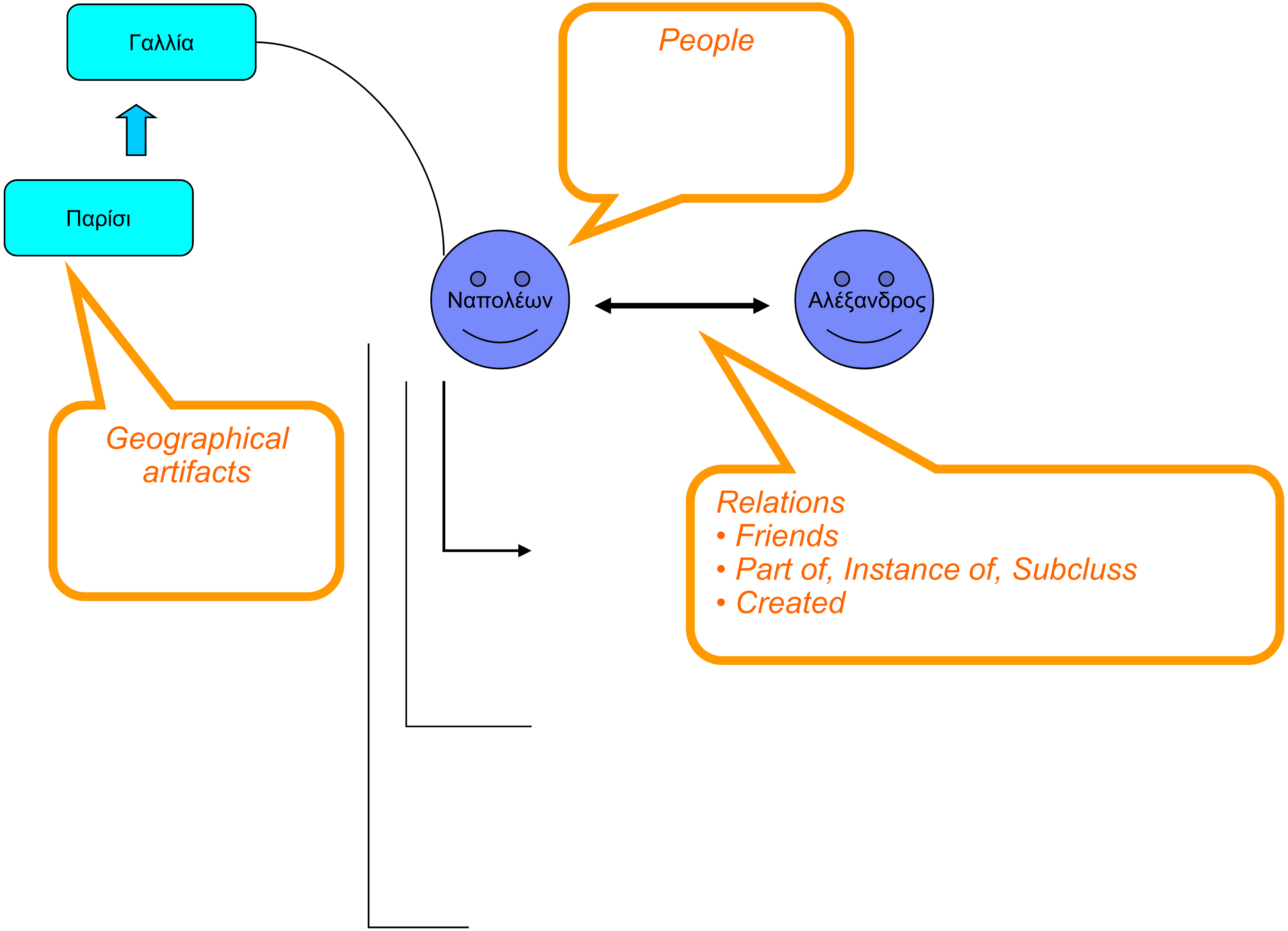
© 2010 Alexander Troussov

How to model the social context?



Log-files of Techno-Social systems (like Facebook or IBM’s Lotus Connections) keep track about *who did what*.
Triples could be aggregated into a network.

© 2010 Alexander Troussov

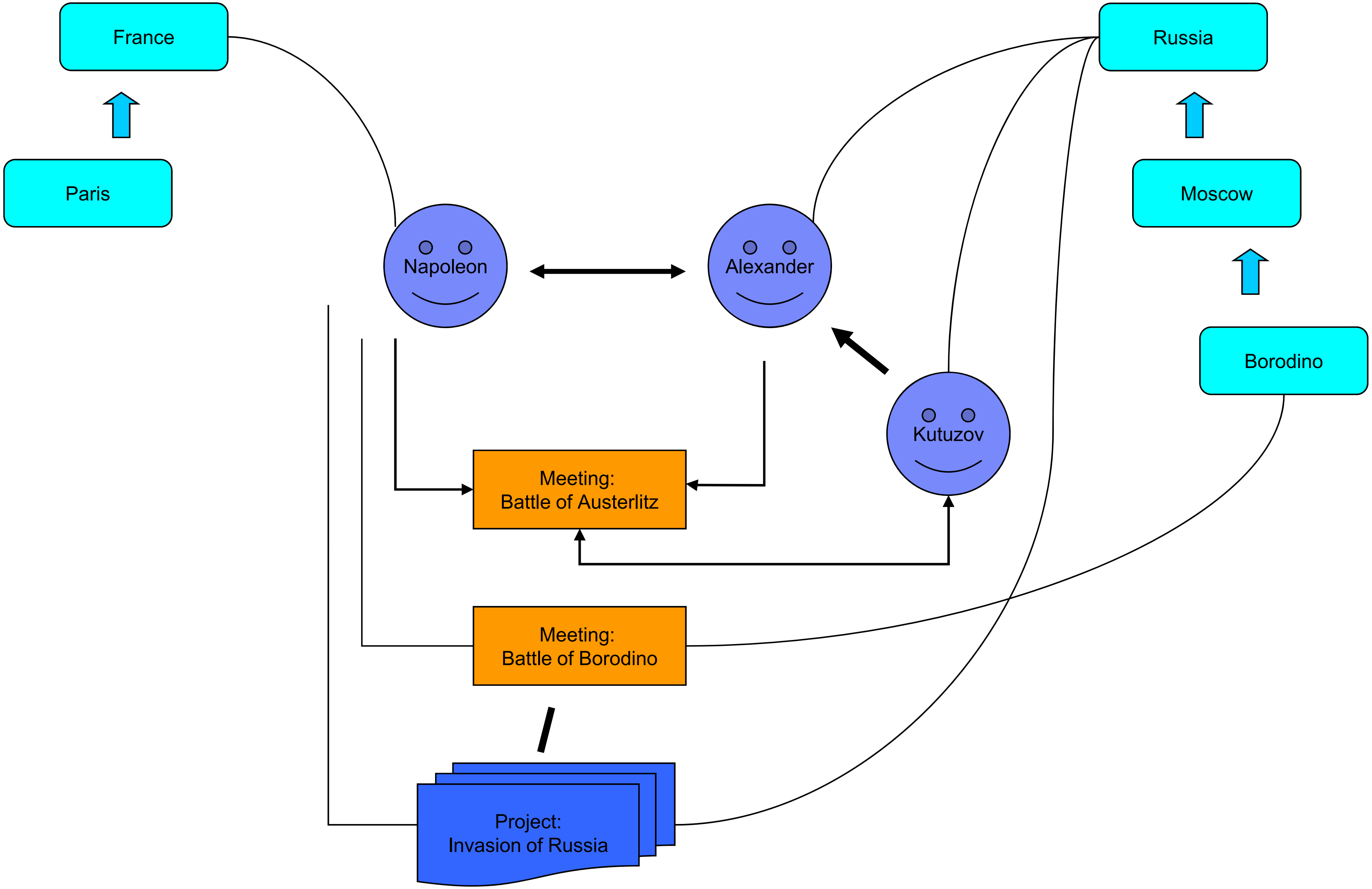


People

Geographical artifacts

Relations
- Friends
- Part of, Instance of, Subcluss
- Created

© 2010 Alexander Troussov

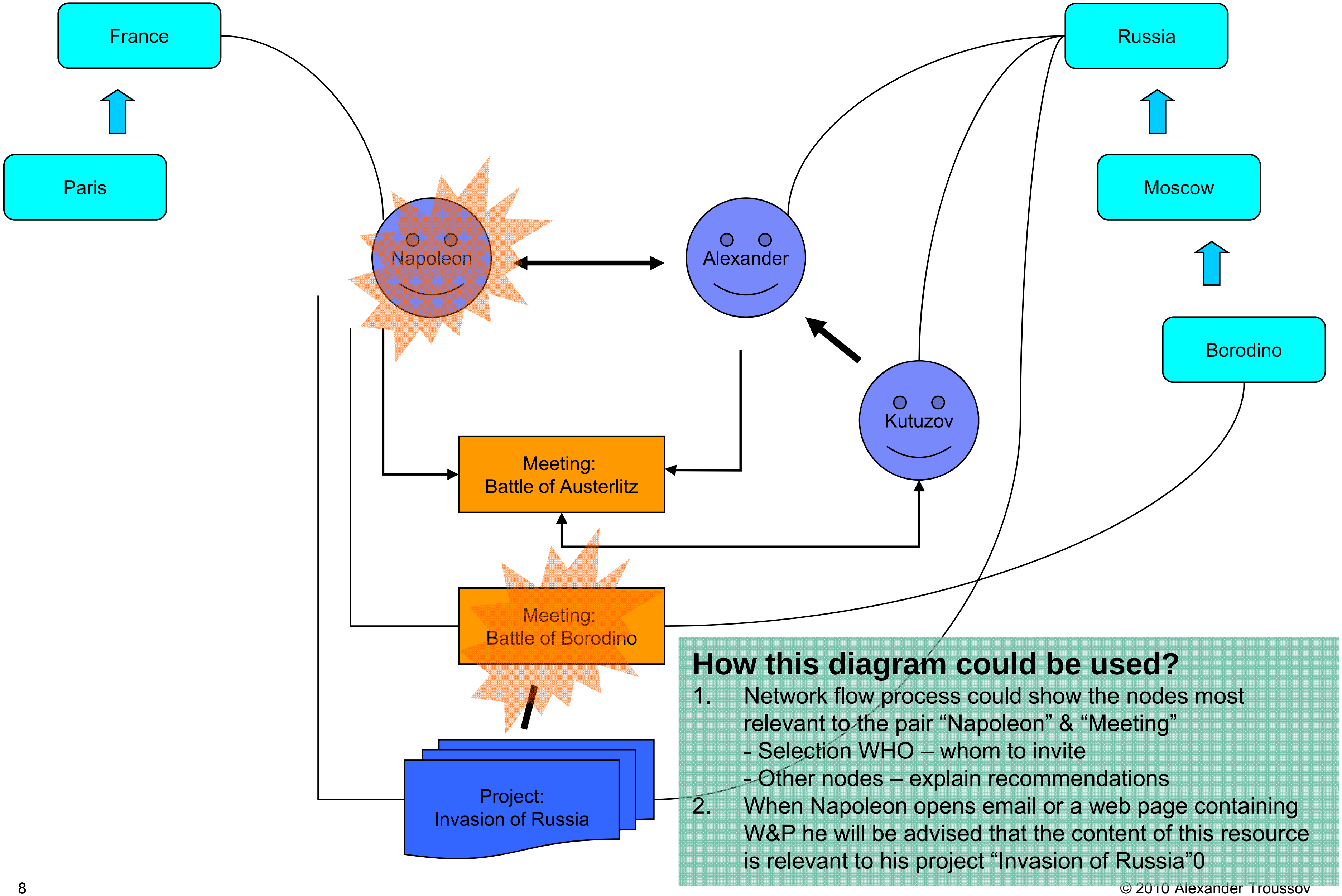


© 2010 Alexander Troussov

Diagram on the previous slide …

- What it represents ?
- How it can be used ?

© 2010 Alexander Troussov



How this diagram could be used?

1. Network flow process could show the nodes most relevant to the pair “Napoleon” & “Meeting”
   - Selection WHO – whom to invite
   - Other nodes – explain recommendations
2. When Napoleon opens email or a web page containing W&P he will be advised that the content of this resource is relevant to his project “Invasion of Russia”0

© 2010 Alexander Troussov

Diagram on the previous slide … What it represents?

- Data from Facebook, data from Napoleon’s Lotus Notes calendar, structure of a Wiki, network of collocations or relations between the entities in W&P, …
  - The proliferation of Web 2.0 and Enterprise 2.0 technologies has lead to the emergence of massive networks connecting people and various digital artifacts. These networks can be treated as a “weak” knowledge, which nevertheless might be used recommendations and even for such traditional applications as knowledge-based text processing
- Or instantiation of an ontology related to W&P by Leo Tolstoy
  - In which case we would probably know that Napoleon is emperor of France, Paris is the capital (not instantiation of a subclass) of France, etc.
- Ontology provides conceptualization, allow inferencing, but these advantages per se are useless without tedious manual work to encode the rules how to use this additional knowledge. While the knowledge encoded in the topology of the multidimensional network is ready to use provided that methods are tolerant to errors and inconsistencies in data - i.e. the methods are methods of “soft mathematic” – fuzzy inferencing, soft clustering, …

© 2010 Alexander Troussov

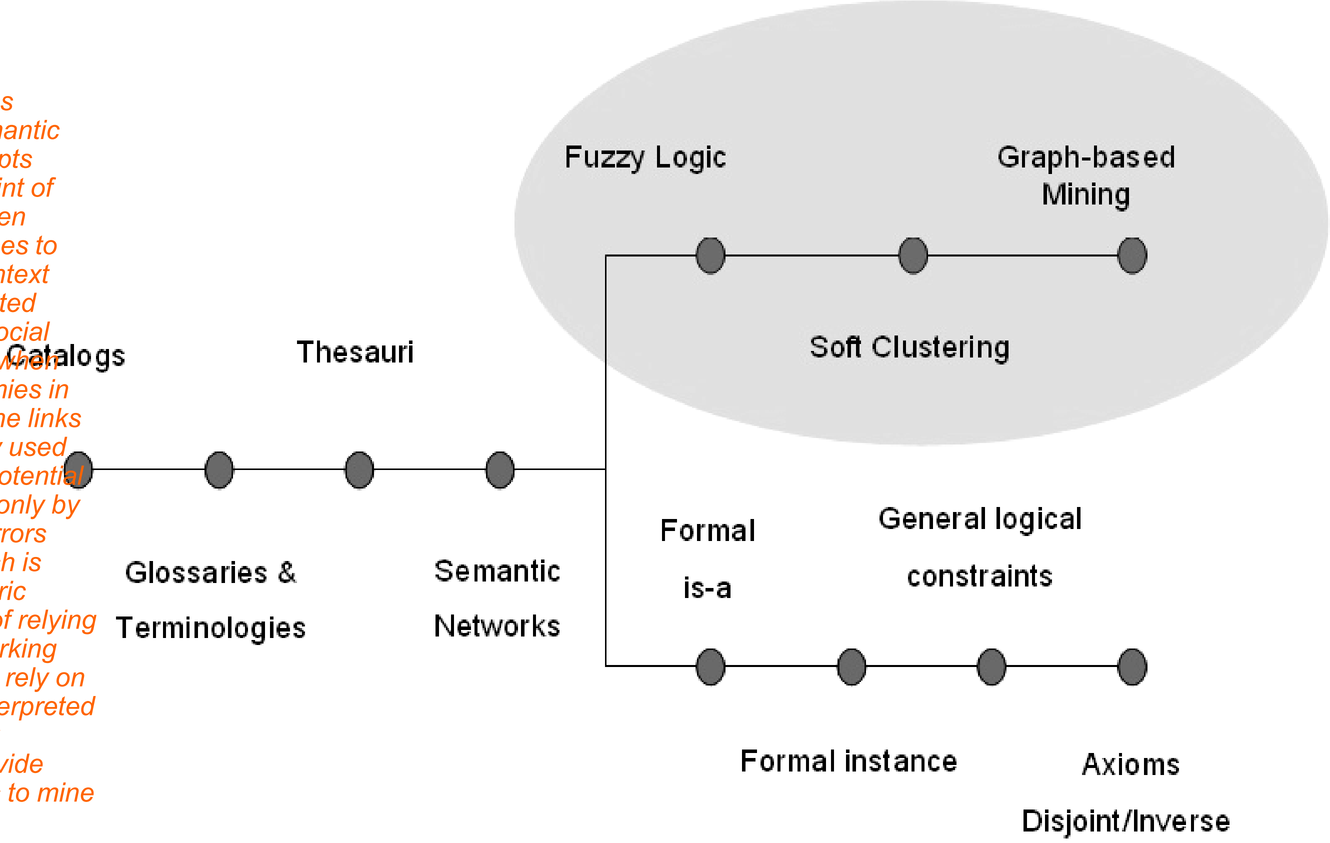
Social Context = Social Knowledge. So what?

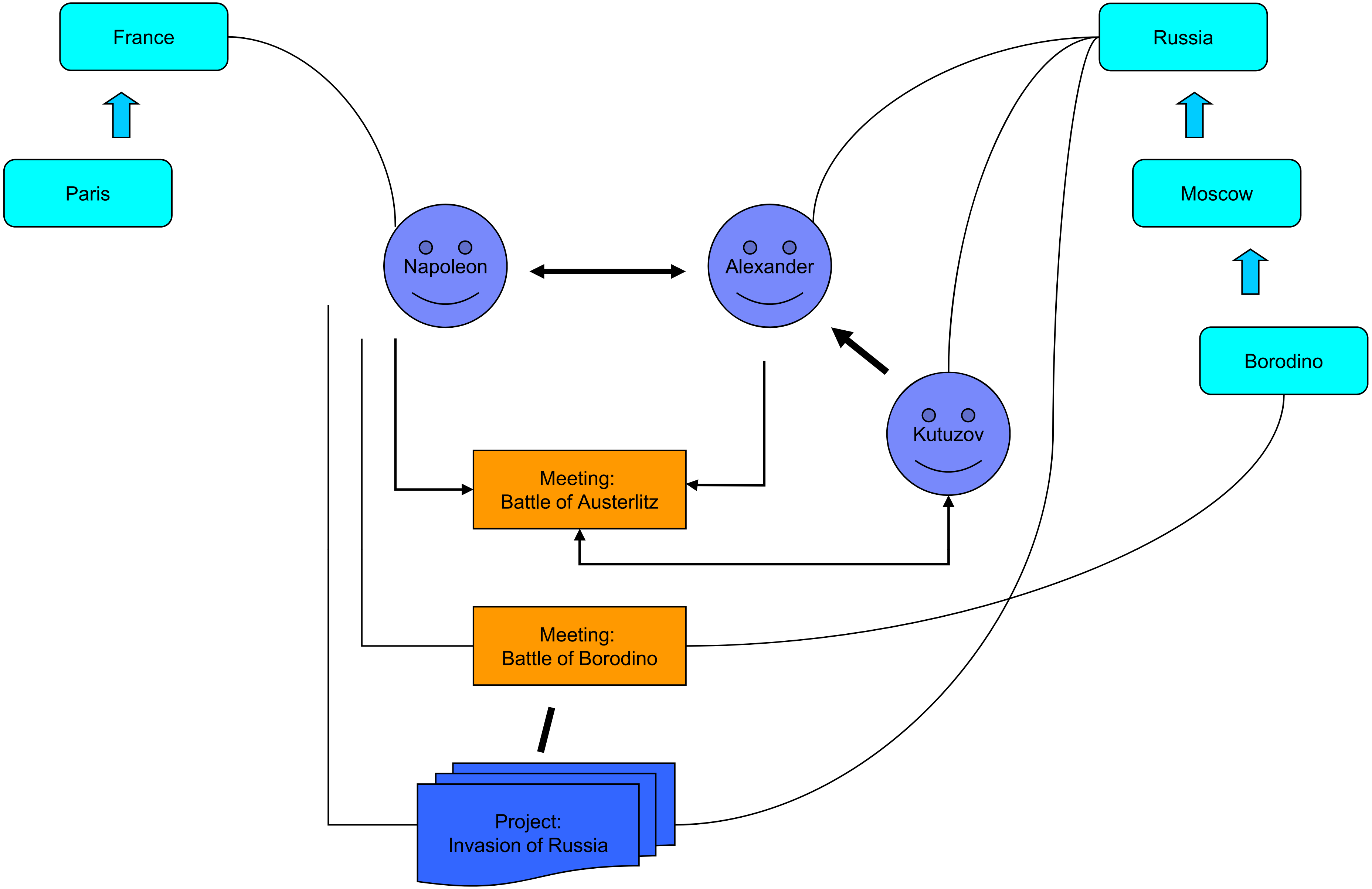
Representing social context as a knowledge allows us to benefit from the past experience of knowledge based applications.

© 2010 Alexander Troussov

For instance, the social context modeled as a network is not much different from semantic networks which are formed from concepts represented in ontologies. And it is possible to use such networks for knowledge based text processing. Representing social context as knowledge allows us to draw experience from such mature R&D area as knowledge-based text processing

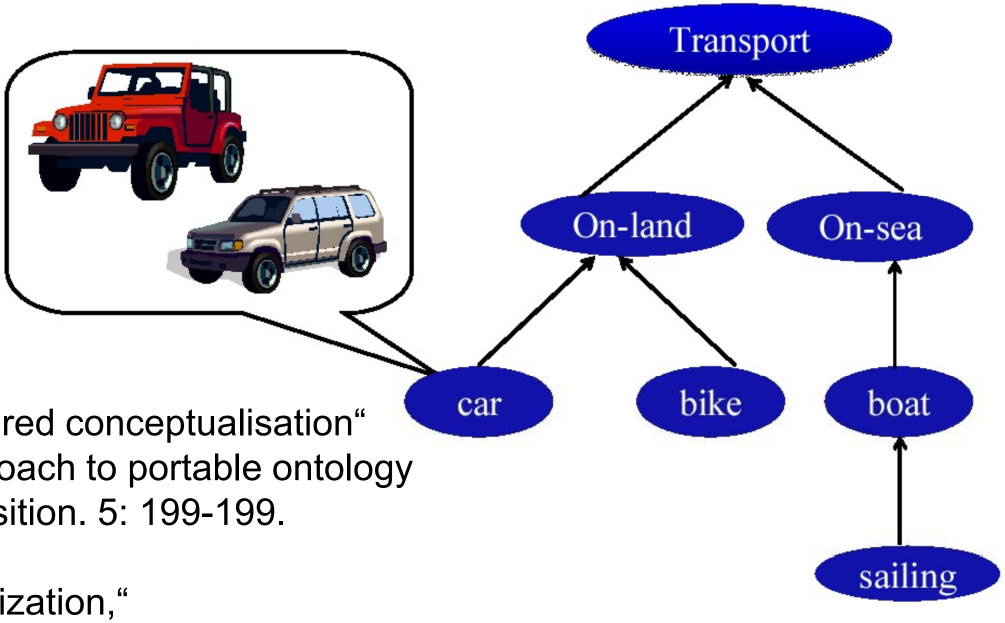
The social context can be considered as knowledge in the same way as the semantic networks which are formed from concepts represented in ontologies. From the point of view of the traditional dichotomy between codification and collaboration approaches to knowledge management, the social context could be considered as bottom-up created social knowledge. As knowledge, the social context is a weaker type of knowledge when contrasted with ontologies and taxonomies in that it lacks proper conceptualisation, the links are usually typed and cannot be readily used for inferencing. Correspondingly, the potential of his knowledge can be fully revealed only by robust methods which are tolerant to errors and incompleteness of knowledge which is endemic in any user created, user centric knowledge system. Therefore instead of relying on the traditional logical methods of working with ontological semantic networks, we rely on graph-based methods which can be interpreted as methods of soft clustering and fuzzy inferencing. Graph-based methods provide clear intuition and elegant mathematics to mine networks.





© 2010 Alexander Troussov

“Strong” Knowledge – Ontologies



- In theory, an ontology is a

  "formal, explicit specification of a shared conceptualisation“
  T. Gruber (1993). "A translation approach to portable ontology specifications". In: Knowledge Acquisition. 5: 199-199.

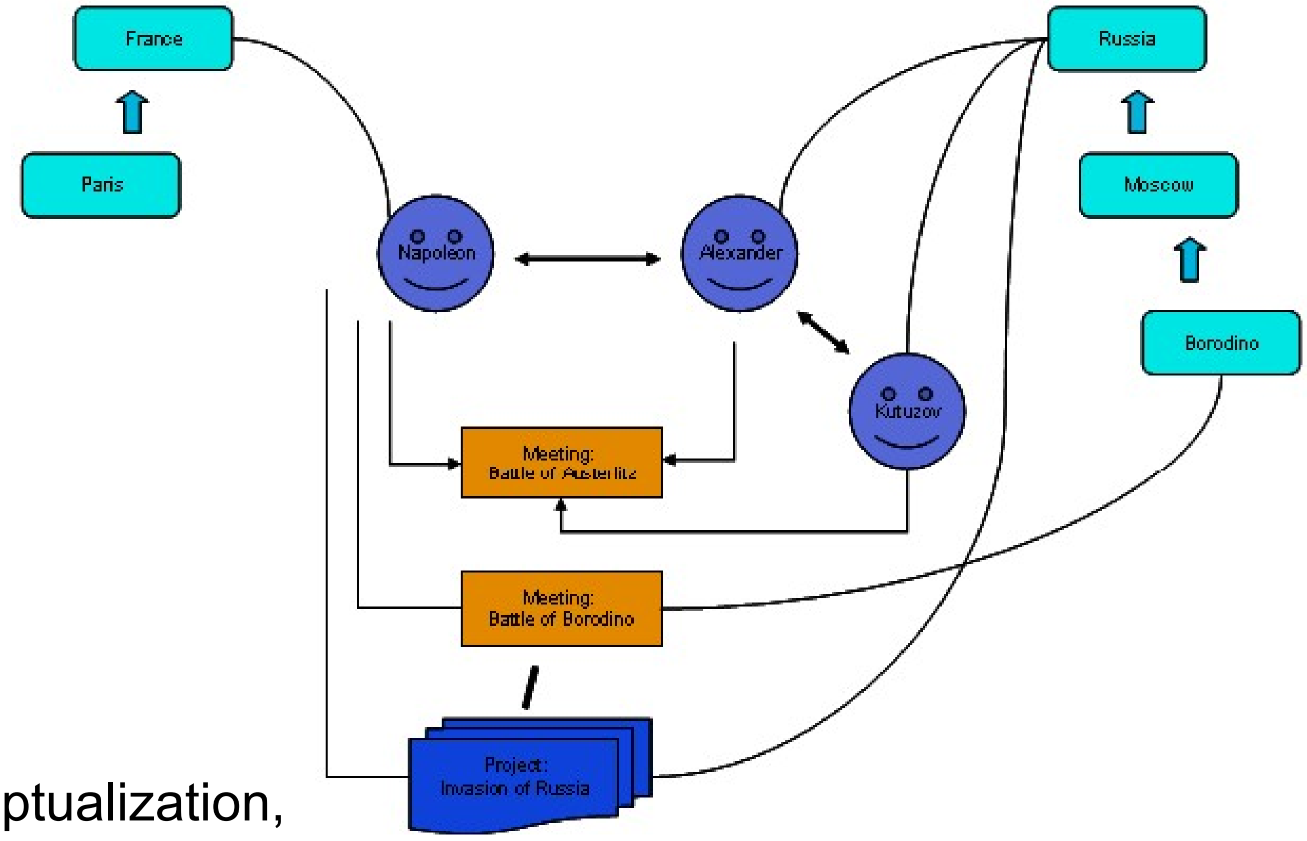
  "explicit specification of a conceptualization,“

  Gruber, T. R., Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies, 43(5-6):907-928, 1995.

- An ontology provides a shared vocabulary, which can be used to model a domain — that is, the type of objects and/or concepts that exist, and their properties and relations

© 2010 Alexander Troussov

“Weak” Knowledge



- No explicit conceptualization,
- Nodes and links are weakly typed.

“Good” Knowledge

- In certain applications manually handcrafted “Strong” knowledge is superior. However, in most modern applications strong knowledge relevant to the domain
    - simply doesn’t exist
    - is outdated
    - and even if available, not always has clear advantages over weak knowledge without additional time-consuming work to encode rules which will benefit from strong knowledge
- While “weak” knowledge is in abundance
    - The primary source - Networked models of socio-technical systems
- and therefore is “Good”

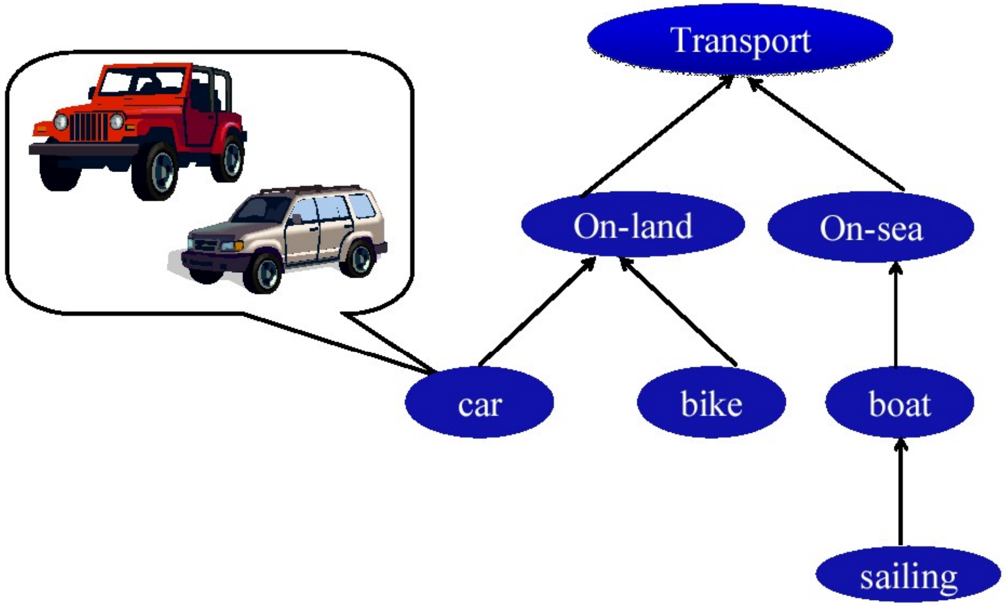
© 2010 Alexander Troussov

Summary and Conclusions of “Social Context as Machine processable knowledge”

- Social Context as Machine-Processable Knowledge
  - Troussov et al. MITACS FP-Nets Workshop on Social Networks, Vancouver, Canada. August 9-13 2010
- Social context could be represented as knowledge (as a weak knowledge as opposed to “strong” knowledge encoded in ontologies)
- Multidimensional networks are capable to model both ontologies and social context
- Social knowledge (primarily network models of techno-social systems) has practical advantages over ontologies
  - Social knowledge is relevant to the domain of applications, is up-to-date, encourages use prior to providing proper structure in full spirit of Web 2.0 business paradigm
- However, social knowledge a “weak knowledge” and its use requires methods of soft mathematics, tolerant to error and inconsistencies in data
- Soft methods actually could provide high-reliable inferencing on networks with high local density based on “weak” knowledge
  - Combinatorial effect leading to a sharp phase transition from uncertainty to certainty

© 2010 Alexander Troussov

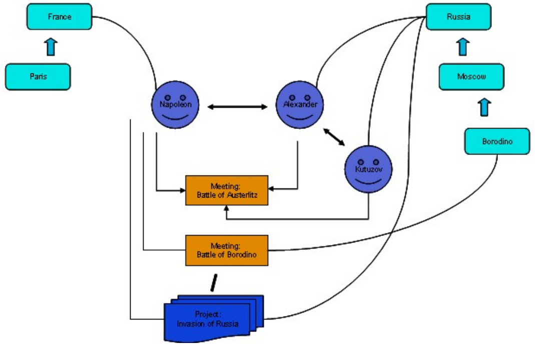
“Strong” Knowledge – Ontologies, Taxonomies



- Ontologies, taxonomies
  - An explicit specification of
    world. Example of real wor
    GeneOntology, Amazon.com Taxonomies
  - Model: a directed graph where a node represents concepts and links represent relations (ISA, …)
  - Having an ontology doesn’t imply having ANY knowledge regarding the real worls
    - For example,

© 2010 Alexander Troussov

“Weak” Knowledge



- No explicit conceptualisation, nodes and links are weakly typed.
  - *some nodes actually represent multiple concepts, like the tag BP could stand for Blue Pages, British Petroleum and a lot more, different nodes should actually be merged into one (remember Wikipedia moderators comments?), etc).*
- A good Example – Folksonomies (and other “crowdsorcing)

© 2010 Alexander Troussov

What Knowledge should we use for recommenders etc?

- It depends …
- and this problem should be properly addressed only in the view of the traditional dichotomy between top-down and bottom-up approaches in knowledge management

© 2010 Alexander Troussov

Good knowledge

- Should be relevant to the domain of the discource
- Know realities
- Be up-to-date

© 2010 Alexander Troussov

What knowledge should we use for text analytics?

- In short:
- In certain applications manually handcrafted “Strong” knowledge is absolutely superiour. However, in most modern applications strong knowledge relevant to the domain simply doesn’t exist
  - is outdated
  - and even if available, has no clear advantages over weak knowledge
- While “weak” knowledge (including who did what to whom on the social site) is in abundance
  – The primary source - Networked models of socio-technical systems
- and therefore is “Good”

© 2010 Alexander Troussov

What knowledge should we use for text analytics?

- As a processable knowledge for understanding the documents embedded into a techno-social system, this social context has advantages over traditional ontologies. The social context is up-to-date knowledge about a subject area or community (like Facebook) which changes rapidly to reflect interests and developments in the area. It is populated with nodes representing the current state of the information and how it relates to processed texts and to the realities of a particular socio-technological system (people, projects, social groups).

© 2010 Alexander Troussov

KNOWLEDGE BASED TEXT PROCESSING

(using network models of weak knowledge)

© 2010 Alexander Troussov

Representing knowledge

- There are a number of options:
  - As *objects*, using the well-accepted techniques of object-oriented analysis and design to capture a model
  - As *clauses* going back to the early days of AI and Lisp
  - As *XML*, using the industry-standard structured mark-up language
  - As *graphs*, making use of the things we know about graph theory
  - As some combination of these
- We are looking for
  - Extensibility
  - Easy of merge heterogeneous information
  - Ease of use
- And our choice is - GRAPHS
  - Knowledge is represented by a multidimensional network which is modeled by a graph
  - And the Nepomuk-Simple will give us examples of extensibility/Easy of merge heterogeneous information/Ease of use
-

Source: Simon Dobson, Declan O'Sullivan

© 2010 Alexander Troussov

Graphs

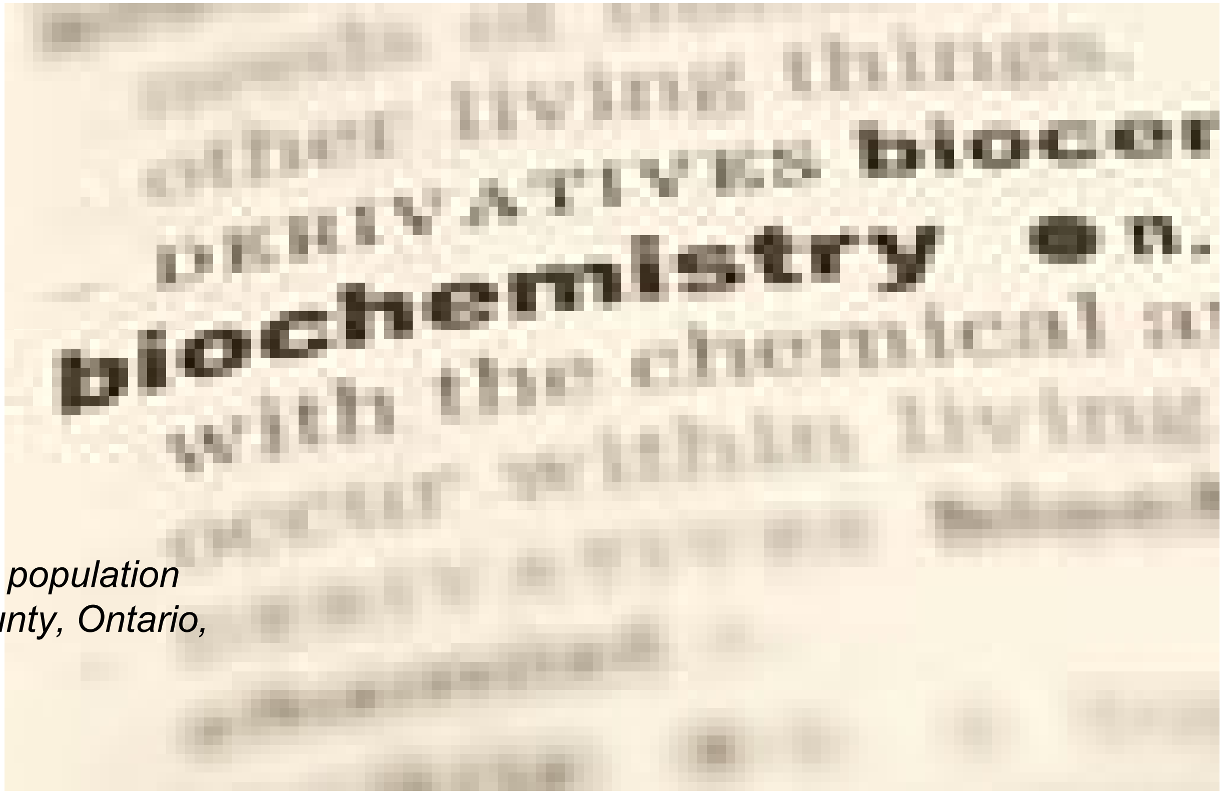
- We can use the nodes of a graph for facts, concepts, people, organisations, etc and the arcs as binary relationships between them
  - Arcs are typically called predicates or relationships in this view
  - The set of arcs intersecting a node tells us the information we know about the fact or entity

Source: Simon Dobson, Declan O'Sullivan

© 2010 Alexander Troussov

Natural Language Understanding is Inferencing

- From computational point of view natural language understanding is inferencing
  - Text which mentions
    Malahide
    is probably about
    Canada *(??)*



Malahide (Canada 2006 Census population 8,828) is a township in Elgin County, Ontario, Canada

© 2010 Alexander Troussov

However …

- Terms are ambiguous, and our knowledge is never “the truth, the whole truth, and nothing but the truth”
  - Malahide, Co. Dublin
  - Malahide is a township in Elgin County, Ontario, Canada.
  - Malahide (Irish: Mullach Íde) is an affluent coastal suburban town, near Dublin city.
  - Malahide United F.C. are a football club from Malahide, County Dublin
  - Malahide Road, Malahide Viaduct, …
  - Paradis Gisenyi Malahide is a hotel in Rwanda
  - Malahide.Net - is an on-line vendor of hot stamp stamping foil.
  - Patrick Malahide - Actor
-

© 2010 Alexander Troussov

Fuzzy Inferencing from Multiple Concepts is a Solution

- One of the successful solutions for this problem in the previous art, is the use of spreading activation (SA) algorithms.
- In our interpretation of this previous art, the success of this methods should be explained as follows:
  - SA effectively provides inferencing from multiple concepts, for instance, the initial seed for the activation propagation starts at two nodes in a geographical taxonomy: *Malahide (Ontario)* and *Malahide (Co. Dublin)* as well as from other concepts mentioned in the text
    - Text which mentions *Malahide* and *Europe* – is a little bit more likely to be about Ireland than about Canada
    - Text which mentions *Malahide* and *Clontarf* – is more likely to be about *Ireland* than about *Canada*
    - *…*
    - Cohesive coherent text which mentions: *Malahide, Mulhuddart, Lansdowne, Clontarf, Donabate* - is almost for sure about *Dublin*
- Such rapid “phase transition” from uncertainty to certainty is similar to the transition related to percolation threshold

© 2010 Alexander Troussov

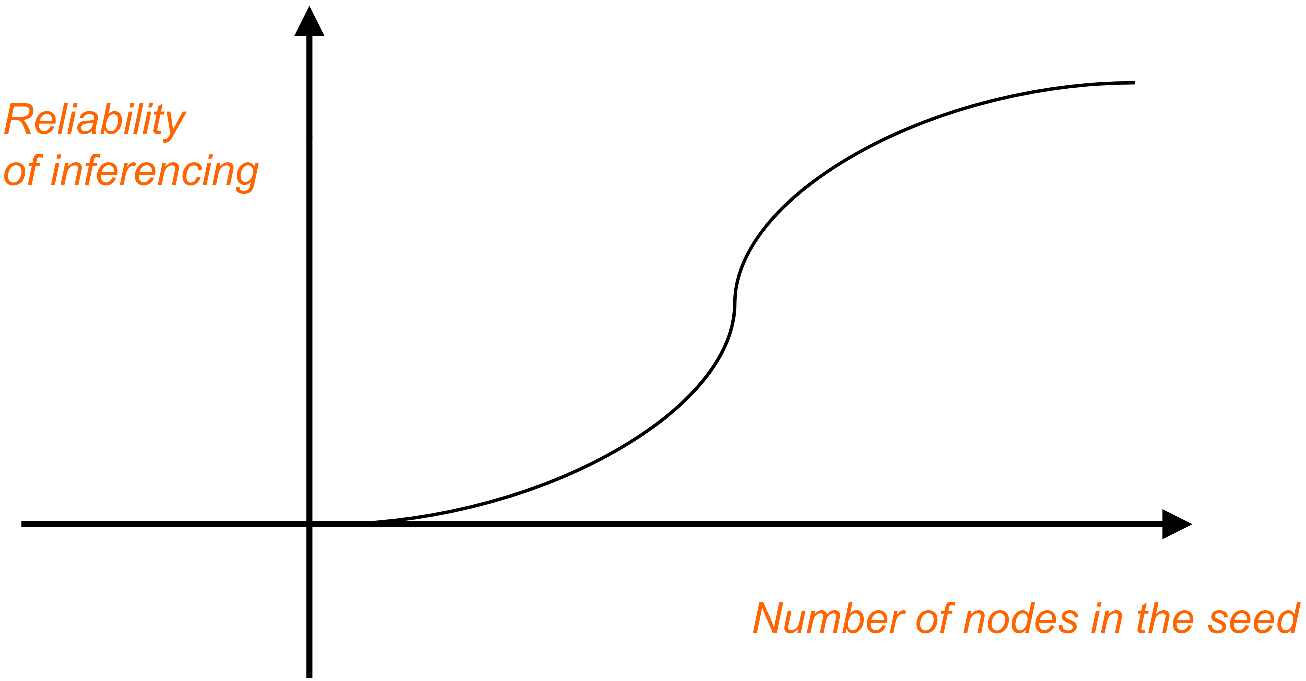
Percolation Threshold

- In physics, chemistry and materials science, percolation concerns the movement and filtering of fluids through porous materials. Percolation threshold is a mathematical term related to percolation theory, which is the formation of long-range connectivity in random systems.
  - Think of a cube of plastic with metal shavings suspended inside. Some of these metal shavings touch; others don't. A high concentration of metal shavings gives you a greater chance of a conductive connection between opposing sides of a cube; lower concentration of metal shavings reduce that chance.
-

© 2010 Alexander Troussov

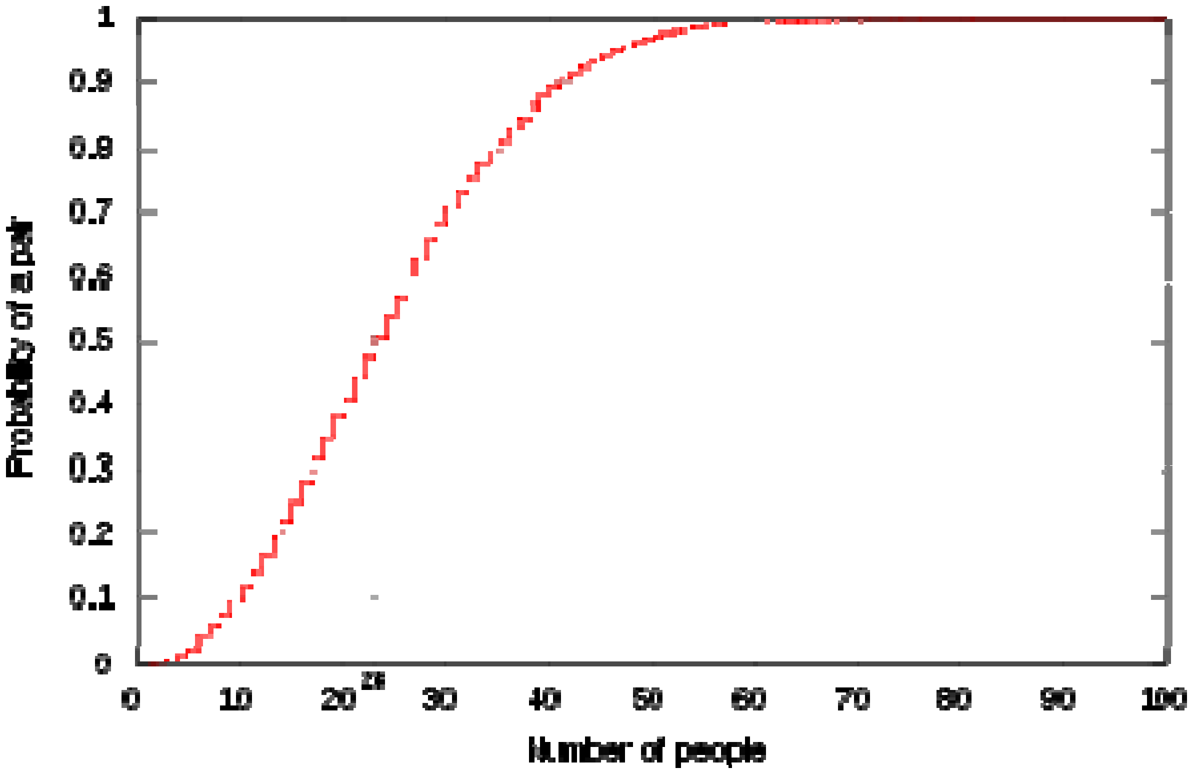
from Uncertainty to Certainty in Inferencing: phase transitions as a function of seed size in analogy to ones in percolation

- In (semantic) networks with high local density
  the reliability of inferencing from a single concept is almost never sufficient,
  reliability could be low when inferencing starts from a small number of seed concepts,
  but inferencing becomes very reliable at some level of the number of the initial seed concepts (which could be explained by combinatorics)



© 2010 Alexander Troussov

and probably could be explained by combinatorics



- A graph showing the approximate probability of at least two people sharing a birthday amongst a certain number of people.

- In probability theory, the birthday problem, or birthday paradox[1] pertains to the probability that in a set of randomly chosen people some pair of them will have the same birthday. By the pigeonhole principle, the probability reaches 100% when the number of people reaches 366 (ignoring February 29 births). But perhaps counter-intuitively, 99% probability is reached with a mere 57 people, and 50% probability with 23 people.

© 2010 Alexander Troussov

Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

RuSSIR

Russian Summer School
in Information Retrieval

Using Social Context for Text Processing

© 2010 Alexander Troussov

In this lecture we’ll show how the social context could be efficiently used for traditional tasks of natural language text processing, such as automated large scale semantic annotation, term disambiguation, search of similar documents, as well as for novel applications such as social recommender systems which aim to alleviate information overload over social media users by presenting the most attractive and relevant content.

© 2010 Alexander Troussov

Bigger context, Individual context

- The current trend in corpus linguistics is for bigger and bigger corpora in order to draw more general analyses.
- In order to provide the type of text analysis needed to drive the development of the social web, we need to look beyond the corpora and documents themselves and draw upon the individual context within which the documents exist.
  - Instead considering how documents in the system relate to each other and also entities (people, tasks, ideas....) outside the scope of the traditional corpus but which have relevance when it comes to analysing the data in the text itself. In addition to word-level, paragraph-level and corpus-level text processing, text analytics on the techno-social level yields a wealth of interesting and useful data and will play increasingly important role in future advances in this area.

© 2010 Alexander Troussov

World of techno-social systems

- We live in an increasingly interconnected world of techno-social systems, in which infrastructures composed of different technological layers are interoperating within the social component that drives their use and development.
  - Nowadays, most of the digital content and metadata is generated within systems like Facebook, Delicious, Twitter, blog and wiki systems, Microsoft SharePoint, and IBM Lotus Connections. These applications have transformed the Web from a mere document collection into a social space where documents are actively exchanged, filtered, organized and discussed.

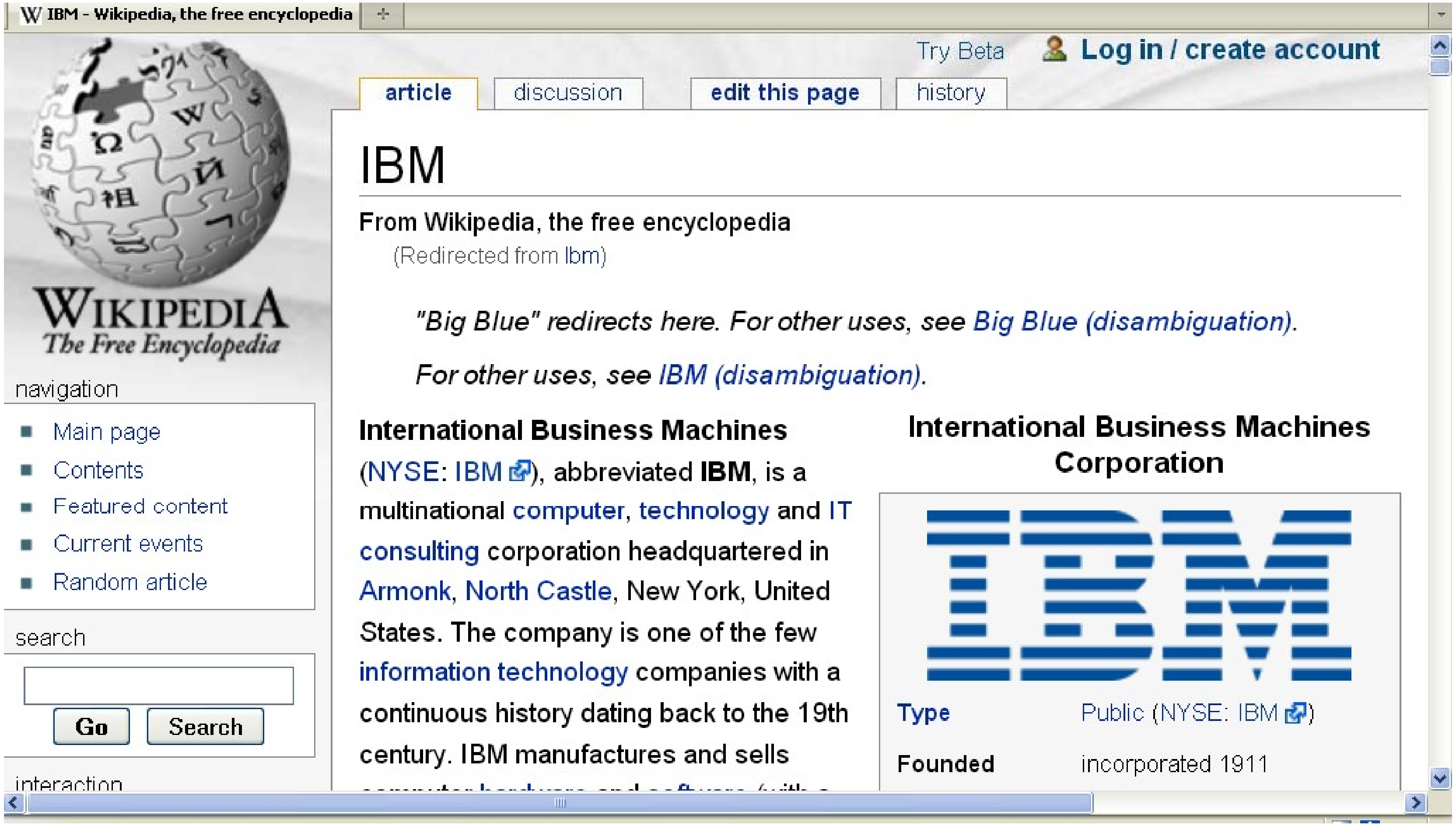
© 2010 Alexander Troussov

World of techno-social systems (cont.)

- In these techno-social systems “everything is deeply intertwingled” using the term coined by the pioneer of the information technologies Ted Nelson: people are connected to other people and to “non-human agents” such as documents, datasets, analytic tools, tags and concepts. And these networks become more and more “multidimensional” providing rich context for processing of embedded natural language texts.

© 2010 Alexander Troussov

Wikipedia – a knowledge repository



IBM

From Wikipedia, the free encyclopedia

(Redirected from Ibm)

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

For other uses, see IBM (disambiguation).

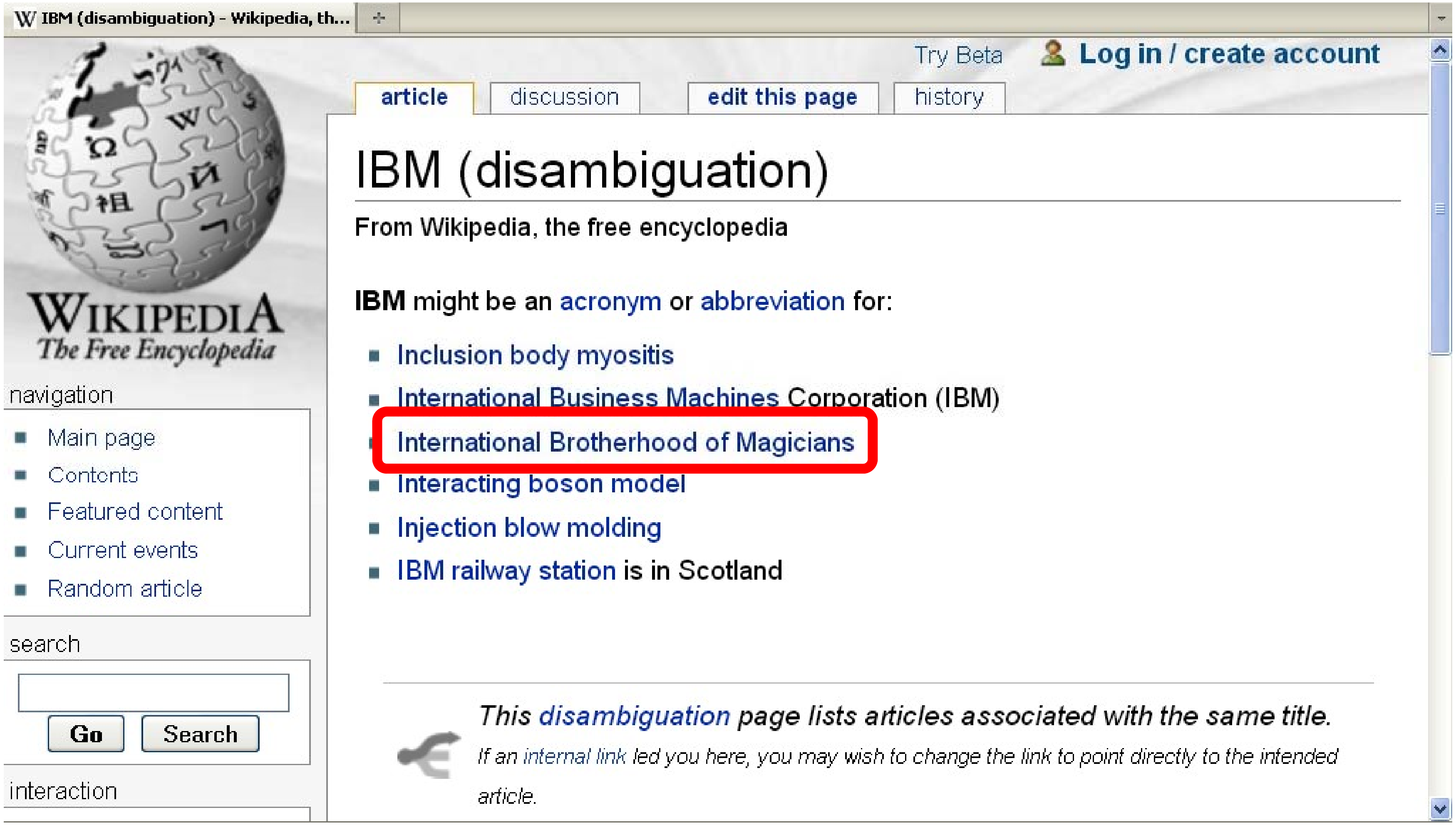
International Business Machines (NYSE: IBM), abbreviated **IBM**, is a multinational computer, technology and IT consulting corporation headquartered in Armonk, North Castle, New York, United States. The company is one of the few information technology companies with a continuous history dating back to the 19th century. IBM manufactures and sells

International Business Machines Corporation

Type	Public (NYSE: IBM)
Founded	incorporated 1911

© 2010 Alexander Troussov

Wikipedia cont. – a lexico-semantic resource



IBM (disambiguation)

From Wikipedia, the free encyclopedia

IBM might be an acronym or abbreviation for:

- Inclusion body myositis
- International Business Machines Corporation (IBM)
- International Brotherhood of Magicians
- Interacting boson model
- Injection blow molding
- IBM railway station is in Scotland

This disambiguation page lists articles associated with the same title.

If an internal link led you here, you may wish to change the link to point directly to the intended article.

© 2010 Alexander Troussov

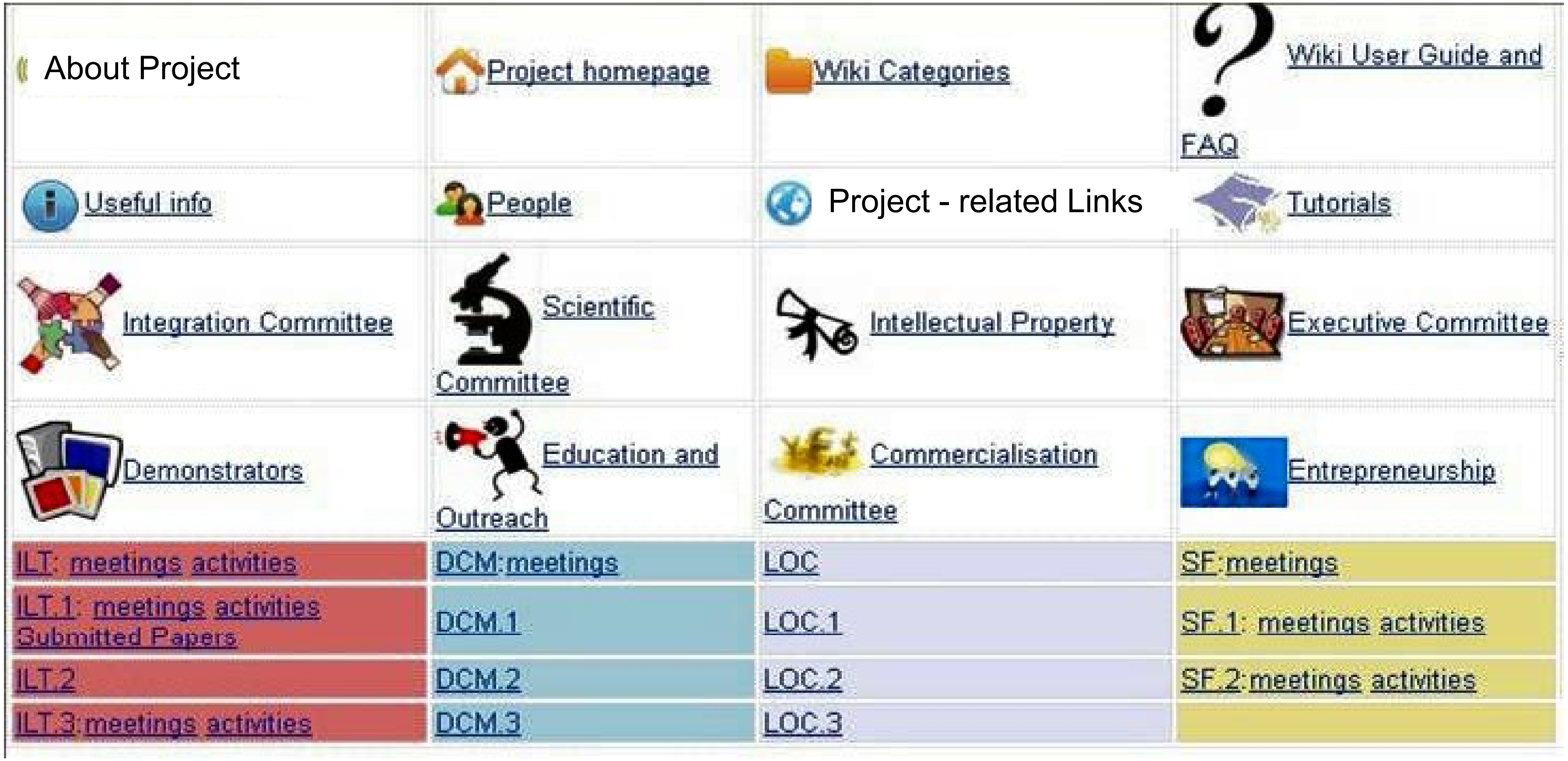
Wikipedia is a techno-social system

- It is based on the technology “wiki” which allows to quickly create and edit pages.
- It is a social systems based on “crowdsourcing”
  - Crowdsourcing is a neologistic compound of Crowd and Outsourcing for the act of taking tasks traditionally performed by an employee or contractor, and outsourcing them to a group of people or community, through an "open call" to a large group of people (a crowd) asking for contributions. The term has become popular with businesses, authors, and journalists as shorthand for the trend of leveraging the mass collaboration enabled by Web 2.0 technologies to achieve business goals.

    Source is reliable – Wikipedia ☺

© 2010 Alexander Troussov

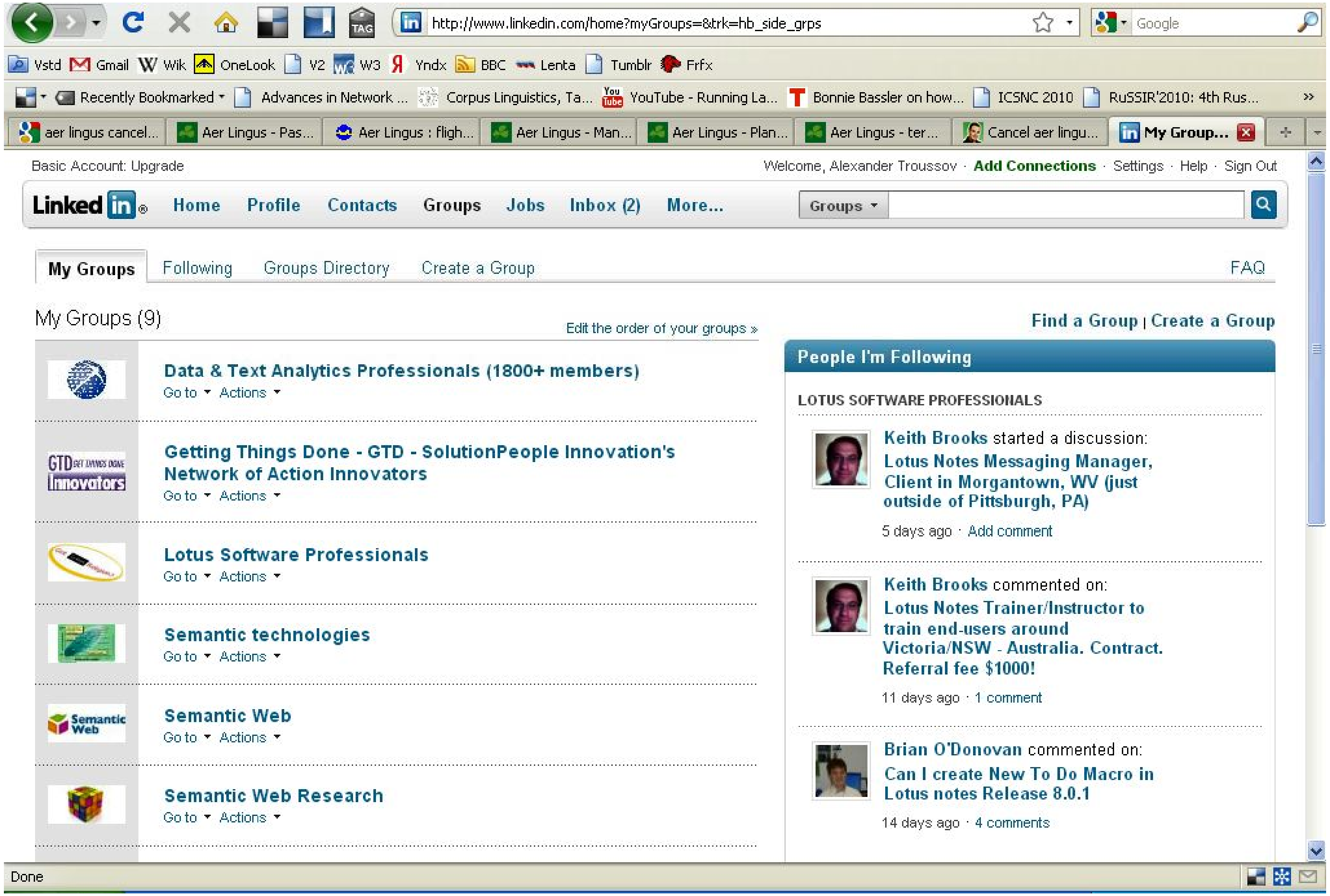
Wikis – popular workplace choice



About Project	Project homepage	Wiki Categories	Wiki User Guide and FAQ
Useful info	People	Project - related Links	Tutorials
Integration Committee	Scientific Committee	Intellectual Property	Executive Committee
Demonstrators	Education and Outreach	Commercialisation Committee	Entrepreneurship
ILT: meetings activities	DCM:meetings	LOC	SF:meetings
ILT.1: meetings activities Submitted Papers	DCM.1	LOC.1	SF.1: meetings activities
ILT.2	DCM.2	LOC.2	SF.2:meetings activities
ILT.3:meetings activities	DCM.3	LOC.3	

© 2010 Alexander Troussov

LinkedIn – Social Net + Groups + …



Basic Account: Upgrade

Welcome, Alexander Troussov · **Add Connections** · Settings · Help · Sign Out

My Groups | Following | Groups Directory | Create a Group | FAQ

My Groups (9)

Edit the order of your groups »

Find a Group | Create a Group

Data & Text Analytics Professionals (1800+ members)
Go to ▾ Actions ▾

Getting Things Done - GTD - SolutionPeople Innovation's Network of Action Innovators
Go to ▾ Actions ▾

Lotus Software Professionals
Go to ▾ Actions ▾

Semantic technologies
Go to ▾ Actions ▾

Semantic Web
Go to ▾ Actions ▾

Semantic Web Research
Go to ▾ Actions ▾

People I'm Following

LOTUS SOFTWARE PROFESSIONALS

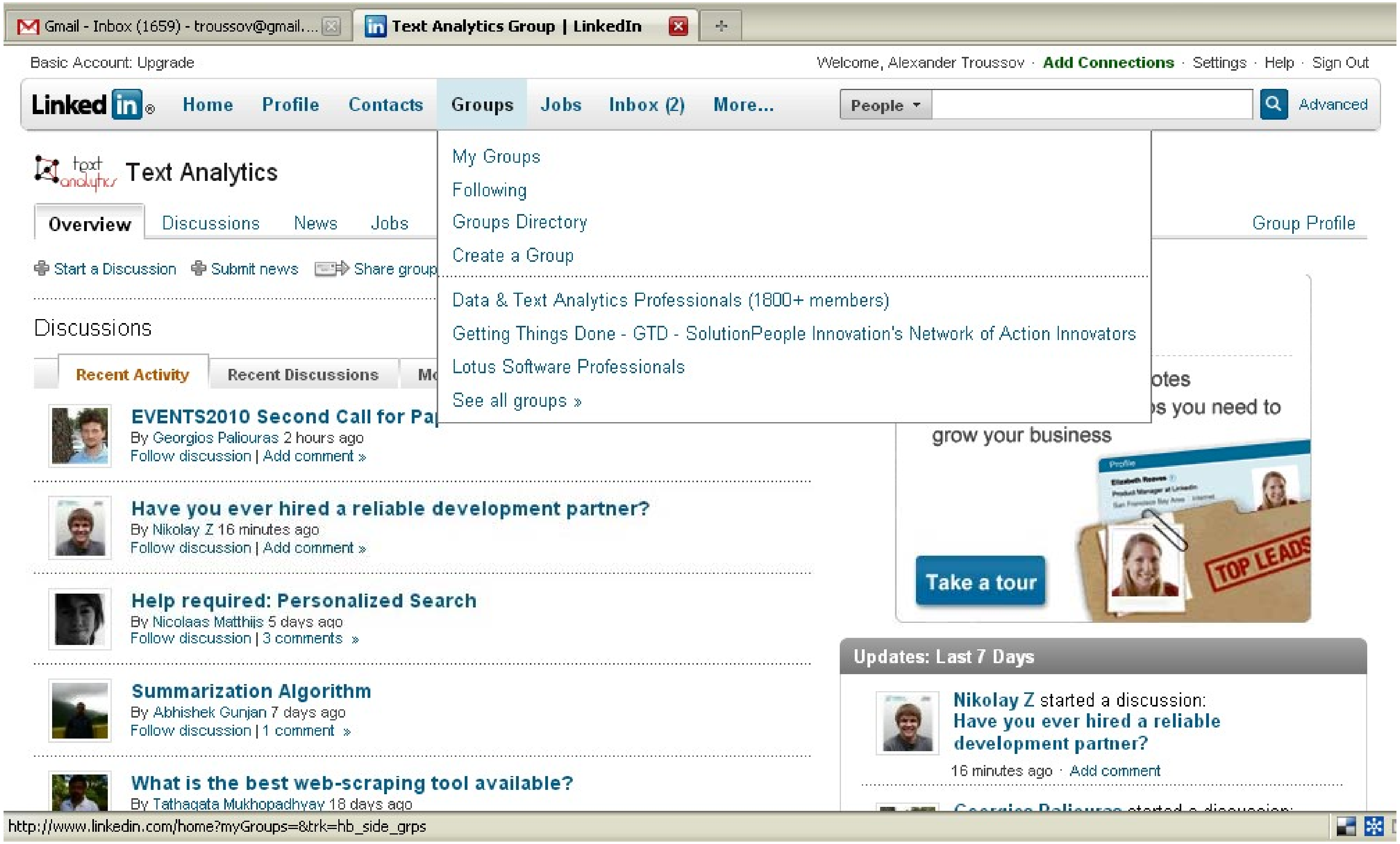
Keith Brooks started a discussion:
Lotus Notes Messaging Manager, Client in Morgantown, WV (just outside of Pittsburgh, PA)
5 days ago · Add comment

Keith Brooks commented on:
Lotus Notes Trainer/Instructor to train end-users around Victoria/NSW - Australia. Contract. Referral fee $1000!
11 days ago · 1 comment

Brian O'Donovan commented on:
Can I create New To Do Macro in Lotus notes Release 8.0.1
14 days ago · 4 comments

© 2010 Alexander Troussov

… + Discussions in Groups + …



Gmail - Inbox (1659) - troussov@gmail.... | Text Analytics Group | LinkedIn

Basic Account: Upgrade

Welcome, Alexander Troussov · Add Connections · Settings · Help · Sign Out

LinkedIn · Home · Profile · Contacts · Groups · Jobs · Inbox (2) · More... · People · Advanced

Text Analytics

Overview · Discussions · News · Jobs · Group Profile

Start a Discussion · Submit news · Share group

- My Groups
- Following
- Groups Directory
- Create a Group
- Data & Text Analytics Professionals (1800+ members)
- Getting Things Done - GTD - SolutionPeople Innovation's Network of Action Innovators
- Lotus Software Professionals
- See all groups »

Discussions

Recent Activity · Recent Discussions · Mo

EVENTS2010 Second Call for Pa
By Georgios Paliouras 2 hours ago
Follow discussion | Add comment »

Have you ever hired a reliable development partner?
By Nikolay Z 16 minutes ago
Follow discussion | Add comment »

Help required: Personalized Search
By Nicolaas Matthijs 5 days ago
Follow discussion | 3 comments »

Summarization Algorithm
By Abhishek Gunjan 7 days ago
Follow discussion | 1 comment »

What is the best web-scraping tool available?
By Tathagata Mukhopadhyay 18 days ago

otes
s you need to
grow your business

Take a tour

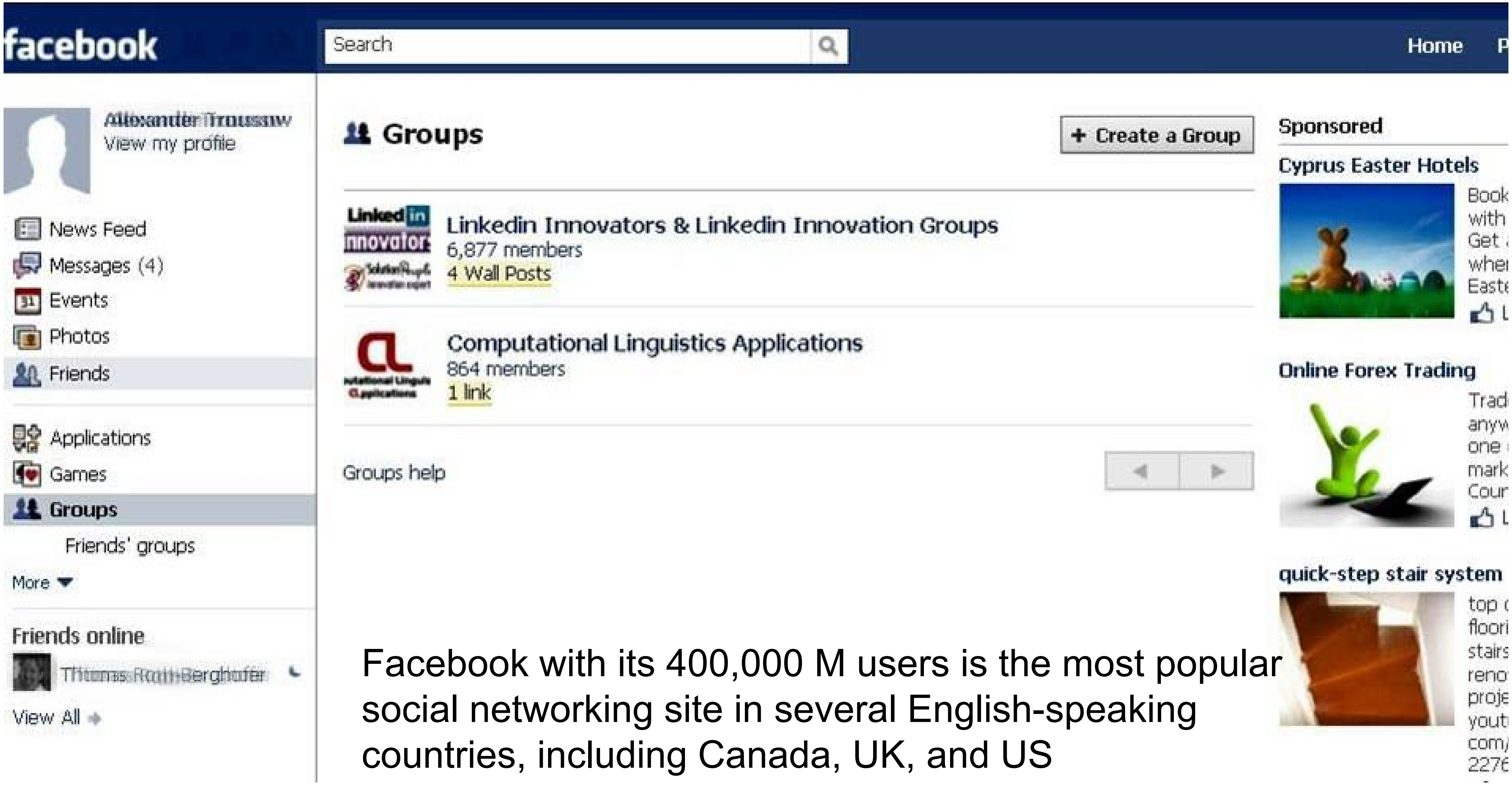
Updates: Last 7 Days

Nikolay Z started a discussion:
Have you ever hired a reliable development partner?
16 minutes ago · Add comment

Georgios Paliouras started a discussion:

http://www.linkedin.com/home?myGroups=&trk=hb_side_grps

© 2010 Alexander Troussov



facebook

Search

Home P

Alexander Troussov
View my profile

- News Feed
- Messages (4)
- Events
- Photos
- Friends
- Applications
- Games
- Groups
- Friends' groups
- More

Friends online

Thomas Roth-Berghofer

View All

Groups

+ Create a Group

Linkedin Innovators & Linkedin Innovation Groups
6,877 members
4 Wall Posts

Computational Linguistics Applications
864 members
1 link

Groups help

Sponsored

Cyprus Easter Hotels
Book with Get wher Easte

Online Forex Trading
Trad anyw one mark Cour

quick-step stair system
top floori stairs reno proje yout com/ 2276

Facebook with its 400,000 M users is the most popular social networking site in several English-speaking countries, including Canada, UK, and US

Facebook directs more online users than Google

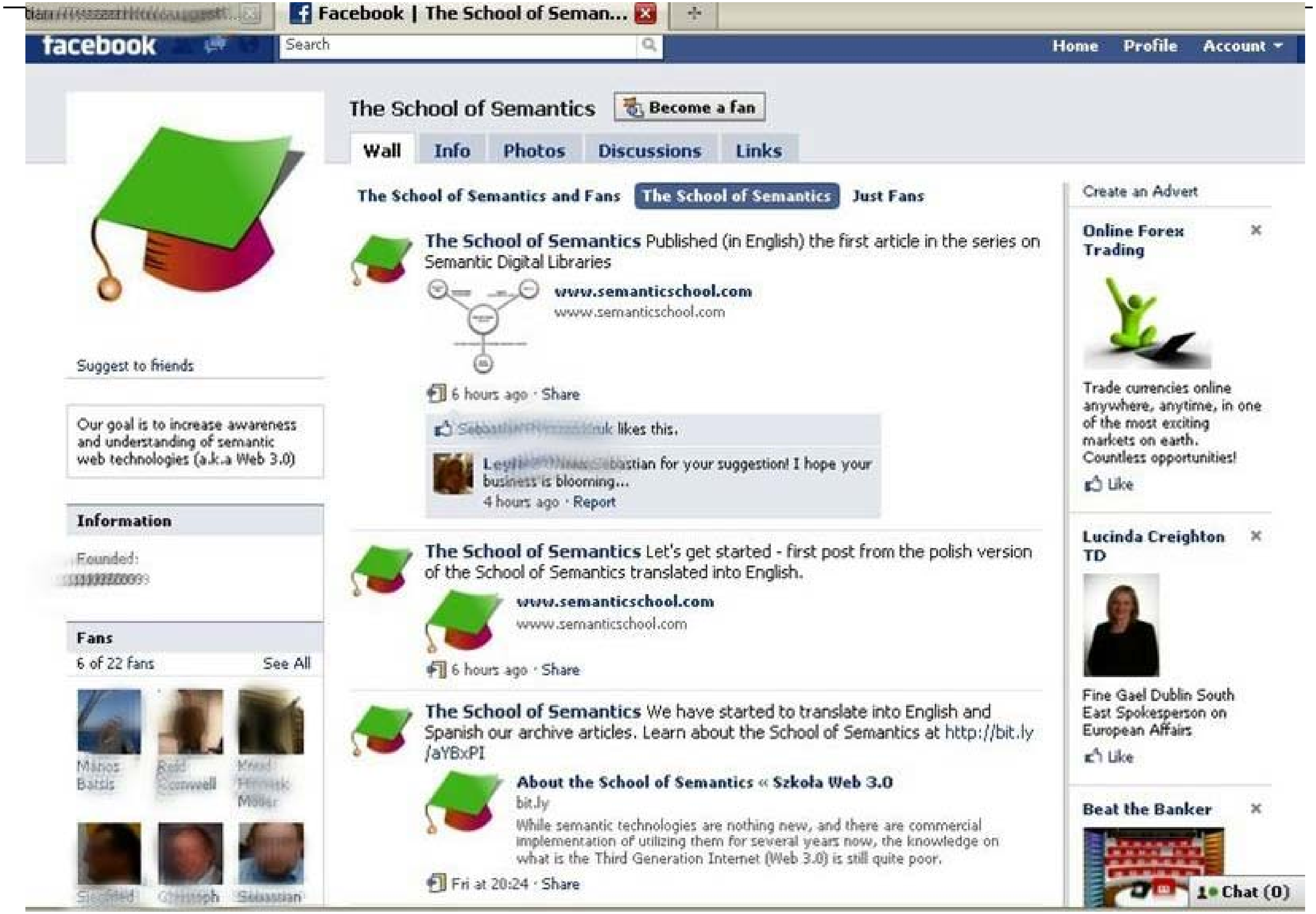
(Benny Evangelista February 15, 2010)

© 2010 Alexander Troussov

Facebook (from Wikipedia)

- Users can create profiles with photos, lists of personal interests, contact information and other personal information. Communicating with friends and other users can be done through private or public messages or a chat feature. Users can also create and join interest and fan groups, some of which are maintained by organizations as a means of advertising.
- Users can add friends and send them messages, and update their personal profiles to notify friends about themselves. Additionally, users can join networks organized by city, workplace, and school or college.
- "How on earth did we stalk our exes, remember our co-workers' birthdays, bug our friends, and play a rousing game of Scrabulous before Facebook?"
- Facebook Notes - a blogging feature that allowed tags and embeddable images.
- By November 3, 2007, seven thousand applications had been developed on the Facebook Platform
- Many new smartphones offer access to the Facebook services either through their web-browsers or applications. Google's Android 2.0 OS automatically includes an official Facebook app.

© 2010 Alexander Troussov



The School of Semantics

Our goal is to increase awareness and understanding of semantic web technologies (a.k.a Web 3.0)

Information

Founded: [redacted]

Fans

6 of 22 fans · See All

[redacted]

The School of Semantics Published (in English) the first article in the series on Semantic Digital Libraries

www.semanticschool.com
www.semanticschool.com

6 hours ago · Share

[redacted]uk likes this.

Le[redacted]bastian for your suggestion! I hope your business is blooming...
4 hours ago · Report

The School of Semantics Let's get started - first post from the polish version of the School of Semantics translated into English.

www.semanticschool.com
www.semanticschool.com

6 hours ago · Share

The School of Semantics We have started to translate into English and Spanish our archive articles. Learn about the School of Semantics at http://bit.ly/aYBxPI

About the School of Semantics « Szkoła Web 3.0
bit.ly
While semantic technologies are nothing new, and there are commercial implementation of utilizing them for several years now, the knowledge on what is the Third Generation Internet (Web 3.0) is still quite poor.

Fri at 20:24 · Share

Online Forex Trading ×

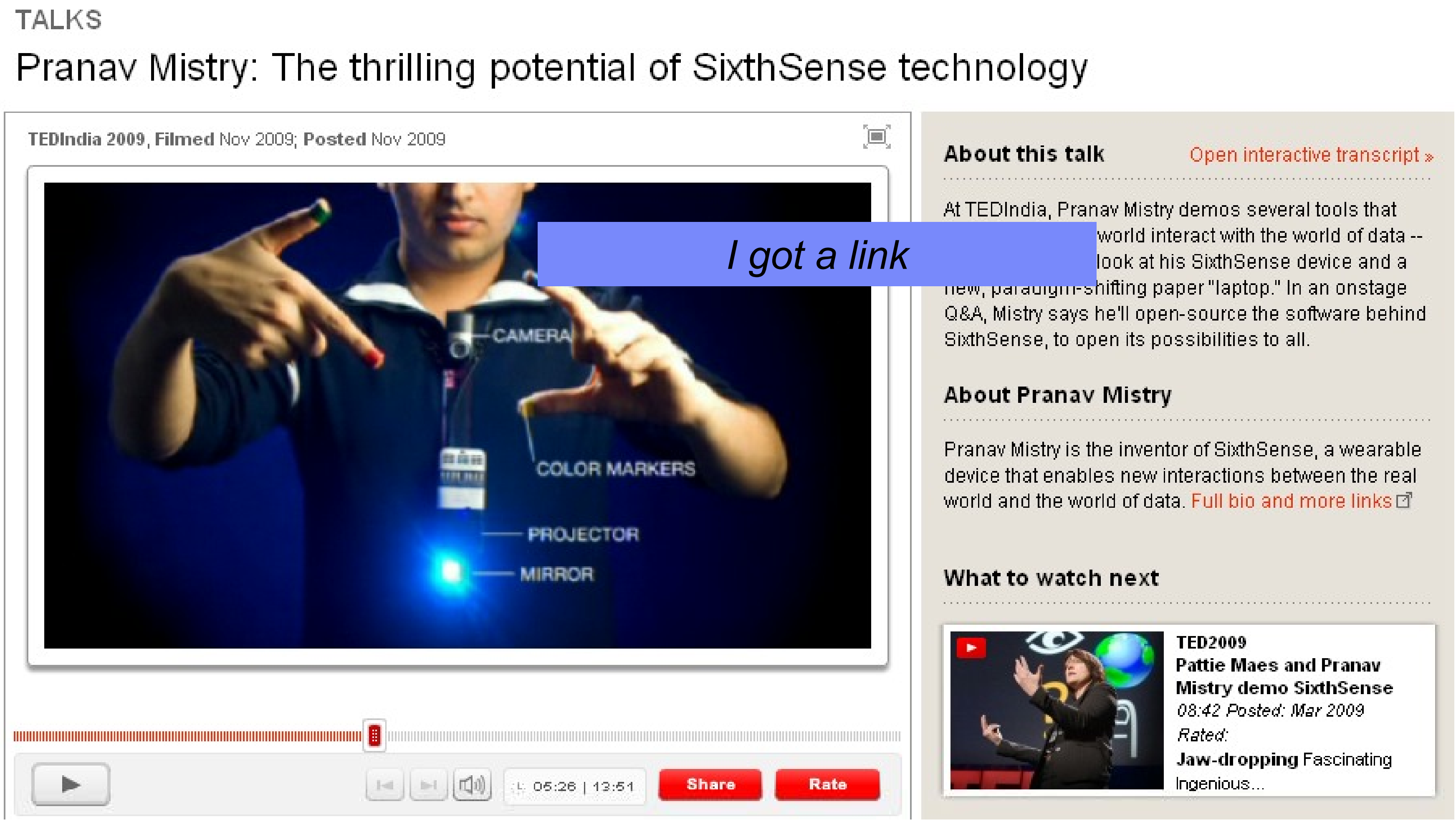
Trade currencies online anywhere, anytime, in one of the most exciting markets on earth. Countless opportunities!
Like

Lucinda Creighton TD ×

Fine Gael Dublin South East Spokesperson on European Affairs
Like

Beat the Banker ×

Sharing, bookmarking, tagging



TALKS

Pranav Mistry: The thrilling potential of SixthSense technology

TEDIndia 2009, **Filmed** Nov 2009; **Posted** Nov 2009

About this talk Open interactive transcript »

At TEDIndia, Pranav Mistry demos several tools that [illegible] world interact with the world of data -- [illegible] look at his SixthSense device and a new, paradigm-shifting paper "laptop." In an onstage Q&A, Mistry says he'll open-source the software behind SixthSense, to open its possibilities to all.

I got a link

About Pranav Mistry

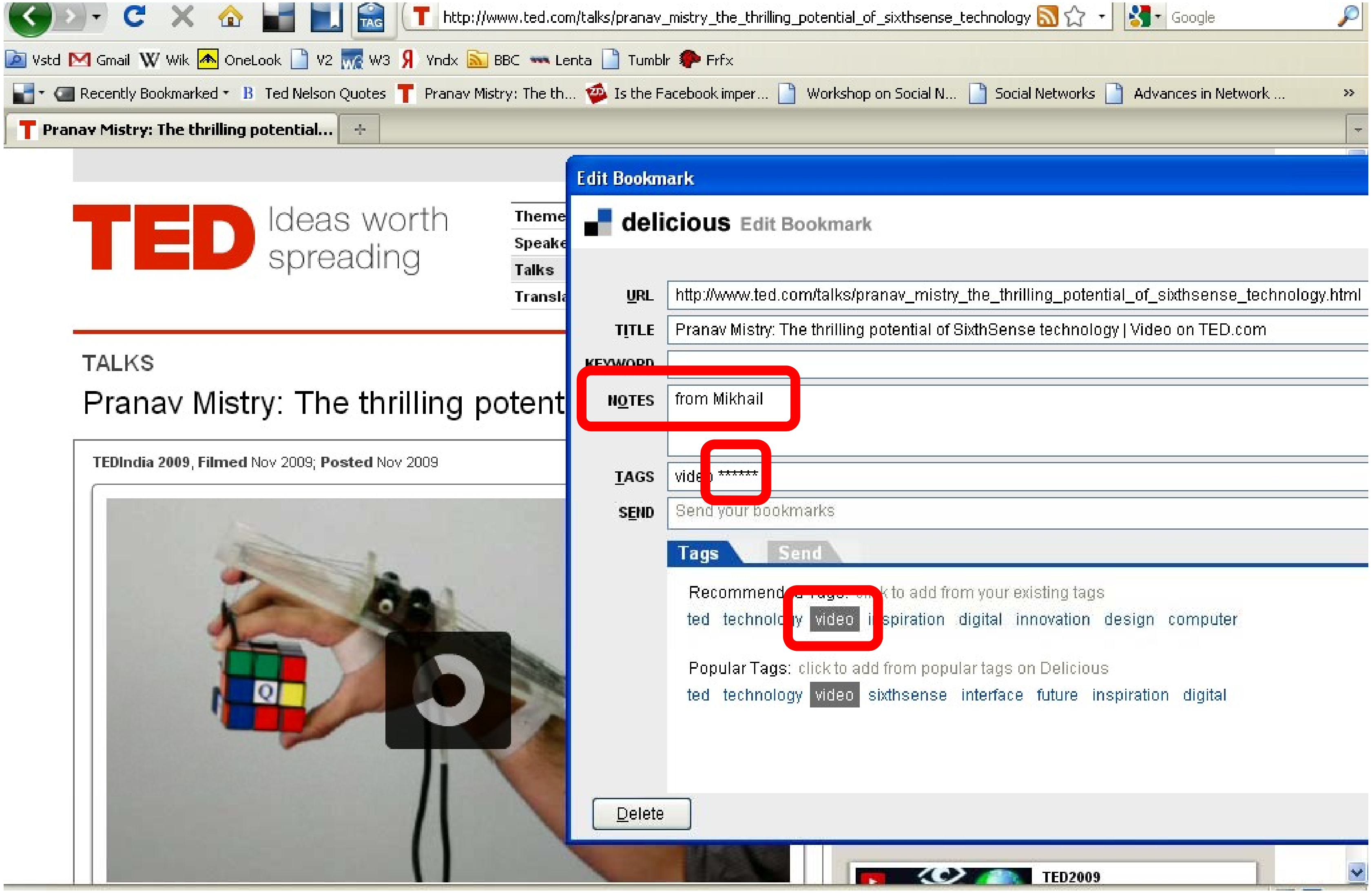
Pranav Mistry is the inventor of SixthSense, a wearable device that enables new interactions between the real world and the world of data. Full bio and more links

What to watch next

TED2009
Pattie Maes and Pranav Mistry demo SixthSense
08:42 Posted: Mar 2009
Rated:
Jaw-dropping Fascinating Ingenious...

© 2010 Alexander Troussov

I bookmarked the link on a collaborative tagging service



Edit Bookmark

delicious Edit Bookmark

URL http://www.ted.com/talks/pranav_mistry_the_thrilling_potential_of_sixthsense_technology.html

TITLE Pranav Mistry: The thrilling potential of SixthSense technology | Video on TED.com

KEYWORD

NOTES from Mikhail

TAGS video ******

SEND Send your bookmarks

Tags Send

Recommended Tags: click to add from your existing tags
ted technology video inspiration digital innovation design computer

Popular Tags: click to add from popular tags on Delicious
ted technology video sixthsense interface future inspiration digital

Delete

© 2010 Alexander Troussov

Tagging – as a part of the Social Context

- Bottom-up approach to semantics to build Folksonomies as the alternative to formal taxonomies (or ontologies)

- Tags reflect all dimensions of human life (see next slide)

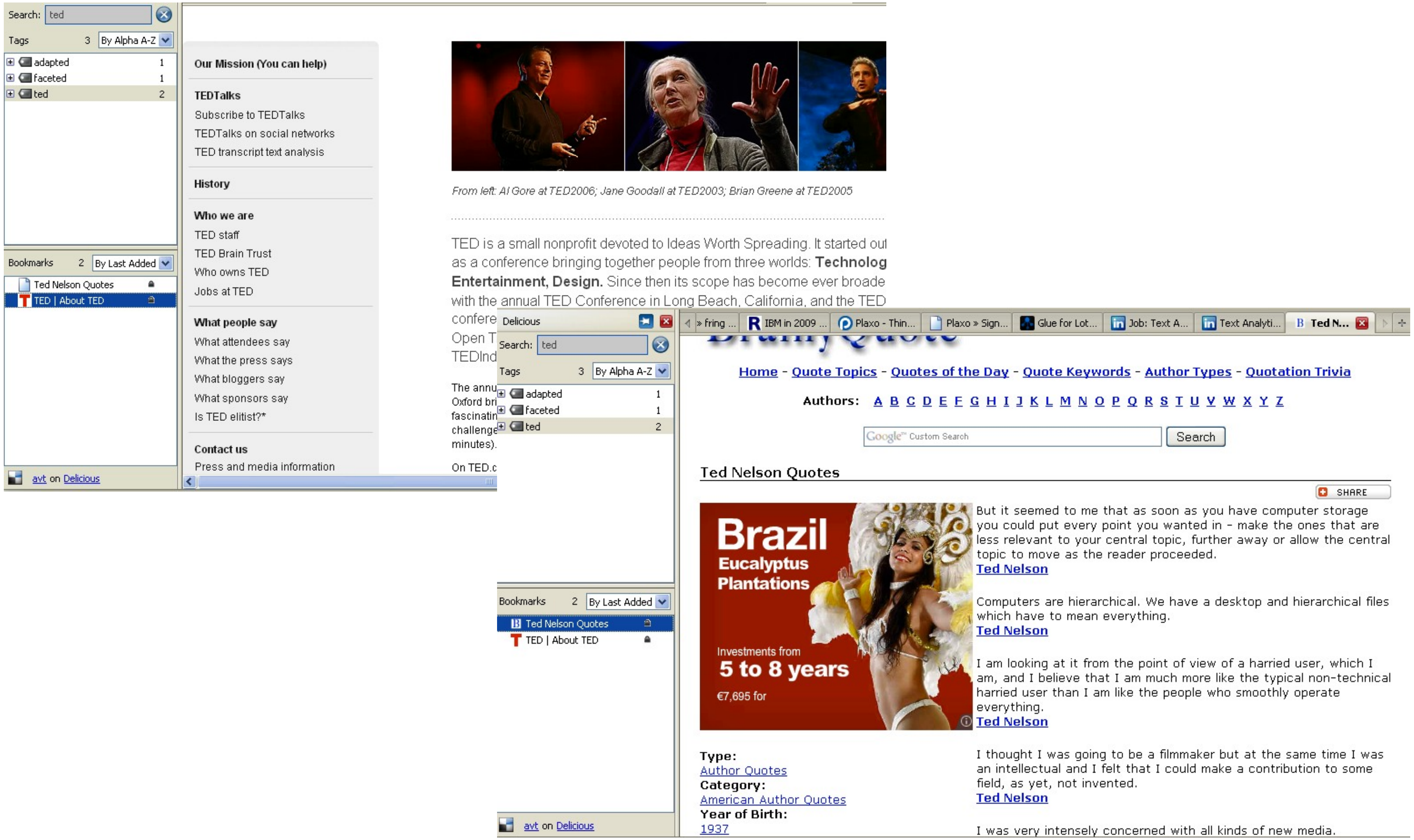
© 2010 Alexander Troussov

Tags reflect dimensions of human life

- Semantics
  - After all, we call ourselves “***homo sapiens***” meaning “Man the Wise”
  - Semantics
    - Semantic web technologies
    - Traditional AI, including Natural Language Understanding
- Social
  - We are social beings as well as individuals / "To live in a society and be free from it is impossible“ / Sometimes we want to be “***homo ludens***” (the “playing man”)
- Activities management
  - “***Homo faber***” (Latin for “Man the Smith” or “Man the Maker”)
  - This is about evocation, “getting things done”, action management, etc

© 2010 Alexander Troussov

Tags are ambiguous



Search: ted

Tags 3 By Alpha A-Z

Tag	Count
adapted	1
faceted	1
ted	2

Bookmarks 2 By Last Added

- Ted Nelson Quotes
- TED | About TED

avt on Delicious

Our Mission (You can help)

TEDTalks

Subscribe to TEDTalks

TEDTalks on social networks

TED transcript text analysis

History

Who we are

TED staff

TED Brain Trust

Who owns TED

Jobs at TED

What people say

What attendees say

What the press says

What bloggers say

What sponsors say

Is TED elitist?*

Contact us

Press and media information

From left: Al Gore at TED2006; Jane Goodall at TED2003; Brian Greene at TED2005

TED is a small nonprofit devoted to Ideas Worth Spreading. It started out as a conference bringing together people from three worlds: **Technolog Entertainment, Design.** Since then its scope has become ever broade with the annual TED Conference in Long Beach, California, and the TED confere Open T TEDInd

The annu Oxford bri fascinatin challenge minutes).

On TED.c

Delicious

Search: ted

Tags 3 By Alpha A-Z

Tag	Count
adapted	1
faceted	1
ted	2

Bookmarks 2 By Last Added

- Ted Nelson Quotes
- TED | About TED

avt on Delicious

Home - Quote Topics - Quotes of the Day - Quote Keywords - Author Types - Quotation Trivia

Authors: A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Google™ Custom Search | Search

Ted Nelson Quotes

SHARE

But it seemed to me that as soon as you have computer storage you could put every point you wanted in - make the ones that are less relevant to your central topic, further away or allow the central topic to move as the reader proceeded.
Ted Nelson

Computers are hierarchical. We have a desktop and hierarchical files which have to mean everything.
Ted Nelson

I am looking at it from the point of view of a harried user, which I am, and I believe that I am much more like the typical non-technical harried user than I am like the people who smoothly operate everything.
Ted Nelson

I thought I was going to be a filmmaker but at the same time I was an intellectual and I felt that I could make a contribution to some field, as yet, not invented.
Ted Nelson

I was very intensely concerned with all kinds of new media.

Type:
Author Quotes
Category:
American Author Quotes
Year of Birth:
1937

© 2010 Alexander Troussov

Social Software

- **Social services are new way of collaboration**
- **Vox populi**
  - Is the Facebook imperative really so great for Corporate America?
  - March 1st, 2010 Posted by Larry Dignan http://blogs.zdnet.com/BTL/?p=31350&tag=nl.e550
    - The question is why is all enterprise software not like Facebook. On Facebook you don’t waste time searching for the right data going from app to app, the data finds you in real time. That’s what customer is getting, not from outdated collaboration applications like Microsoft Sharepoint and Lotus Notes.
      Sharepoint is owned by a lot of businesses, used by far fewer, and enjoyed practically by none. When was the last time everyone said they really loved using Microsoft Sharepoint or Lotus Notes? The reality is customers want to attract their coworkers who matter to them, the most critical conversations, the apps they depend on, the concept they create and share, all using the new mobile devices that they are carrying around in their hand. They want to collaborate without the cost, complexity and flexibility and overall enterprise dead weight of enterprise software, hardware and data centers.

© 2010 Alexander Troussov

Social Context = Knowledge ?

A New Mathematical Model of Horse Racing

- *Assume, without the loss of generality, that each horse in the horse racing is modelled by a wooden ball of radius R_i.*

= a ball ? ☺

© 2010 Alexander Troussov

What kind of knowledge?

- As a knowledge – “the social context” – is a rather “weak” knowledge as compared to “traditional” ontologies
  - *some nodes actually represent multiple concepts, like the tag BP could stand for Blue Pages, British Petroleum and a lot more, different nodes should actually be merged into one (remember Wikipedia moderators comments?), etc).*
- However, “the social context” is up-to-date knowledge about a universe (like Facebook) which changes every second. Therefore as a resource for processing of messages on Facebook it is better than an ontology
  - which has no concepts of Web, Blog, Wiki, and is not populated with instances of global multinationals (Microsoft, Google), influential people (Barak Abama, Tim Berners-Lee, …),
  - and doesn’t know my friends and colleagues,
  - and doesn’t know concepts within the scope of my interests

© 2010 Alexander Troussov

Weak knowledge ☹

- Social knowledge is a weak knowledge.

- The knowledge encoded in ontologies is the truth *?*
  - *"the truth, the whole truth, and nothing but the truth"*

- *Probably not*
  - *After many years as an expert, I've become more and more uncomfortable about swearing to tell the truth, the whole truth and nothing but the truth, especially when I've looked back on cases in which two experts have said exactly the opposite and one's bound to ask which one was telling the truth? ...*

    Professor Max Sussman, The Expert Witness Institute
  - *The truth is rarely pure and never simple.*

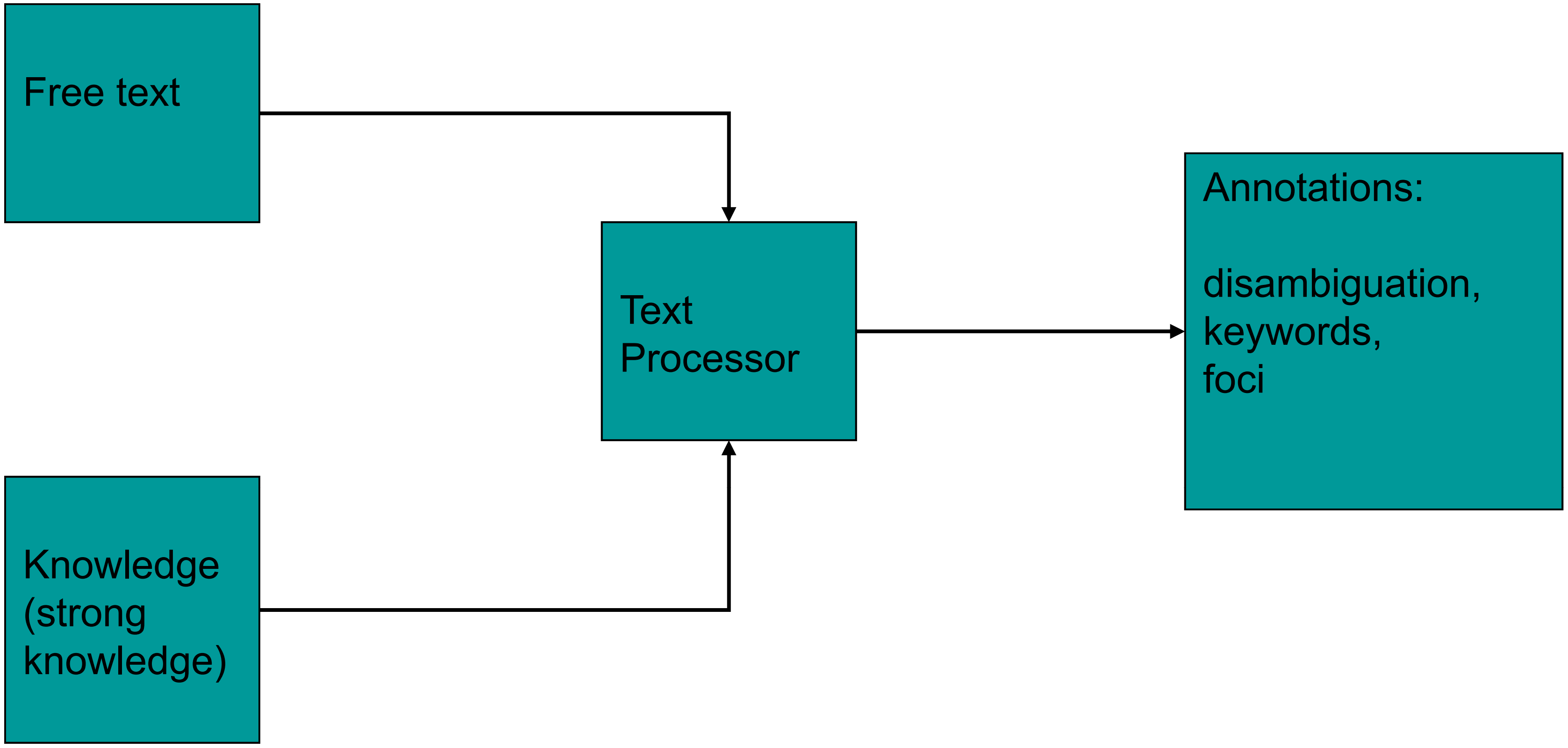
    Oscar Wilde, Irish dramatist, novelist, & poet

© 2010 Alexander Troussov

How such knowledge could be used for text analytics

© 2010 Alexander Troussov

Traditional knowledge-based text processing



Free text

Text Processor

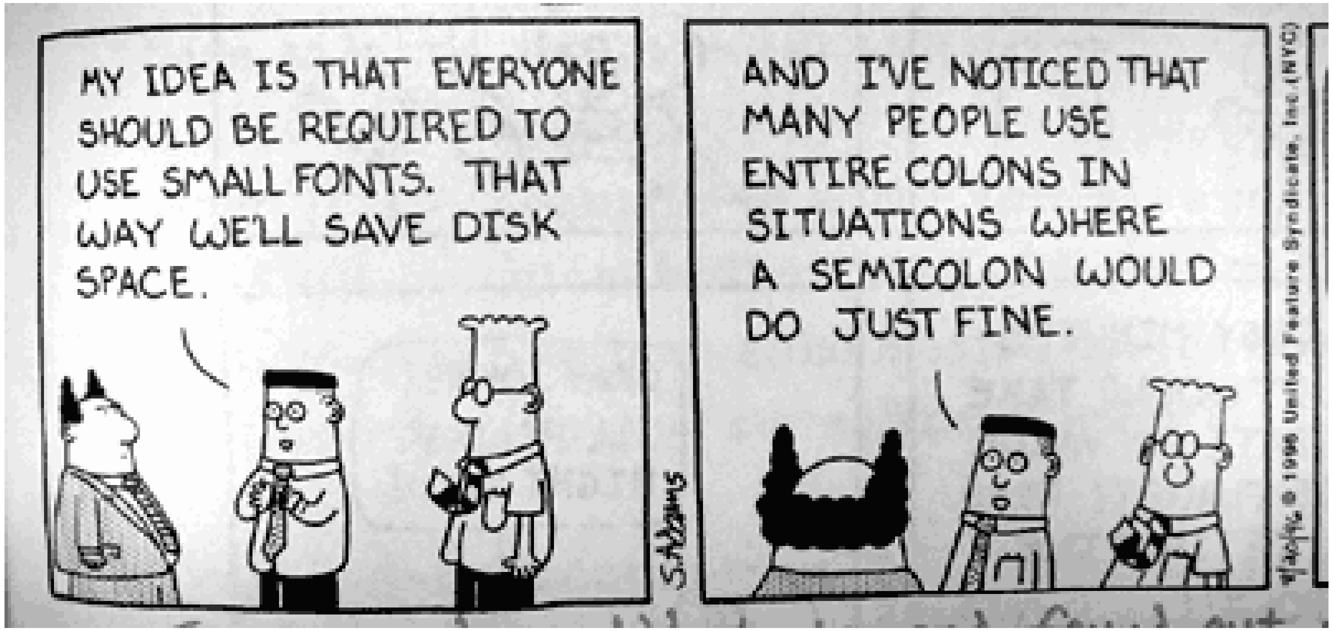
Annotations:

disambiguation, keywords, foci

Knowledge (strong knowledge)

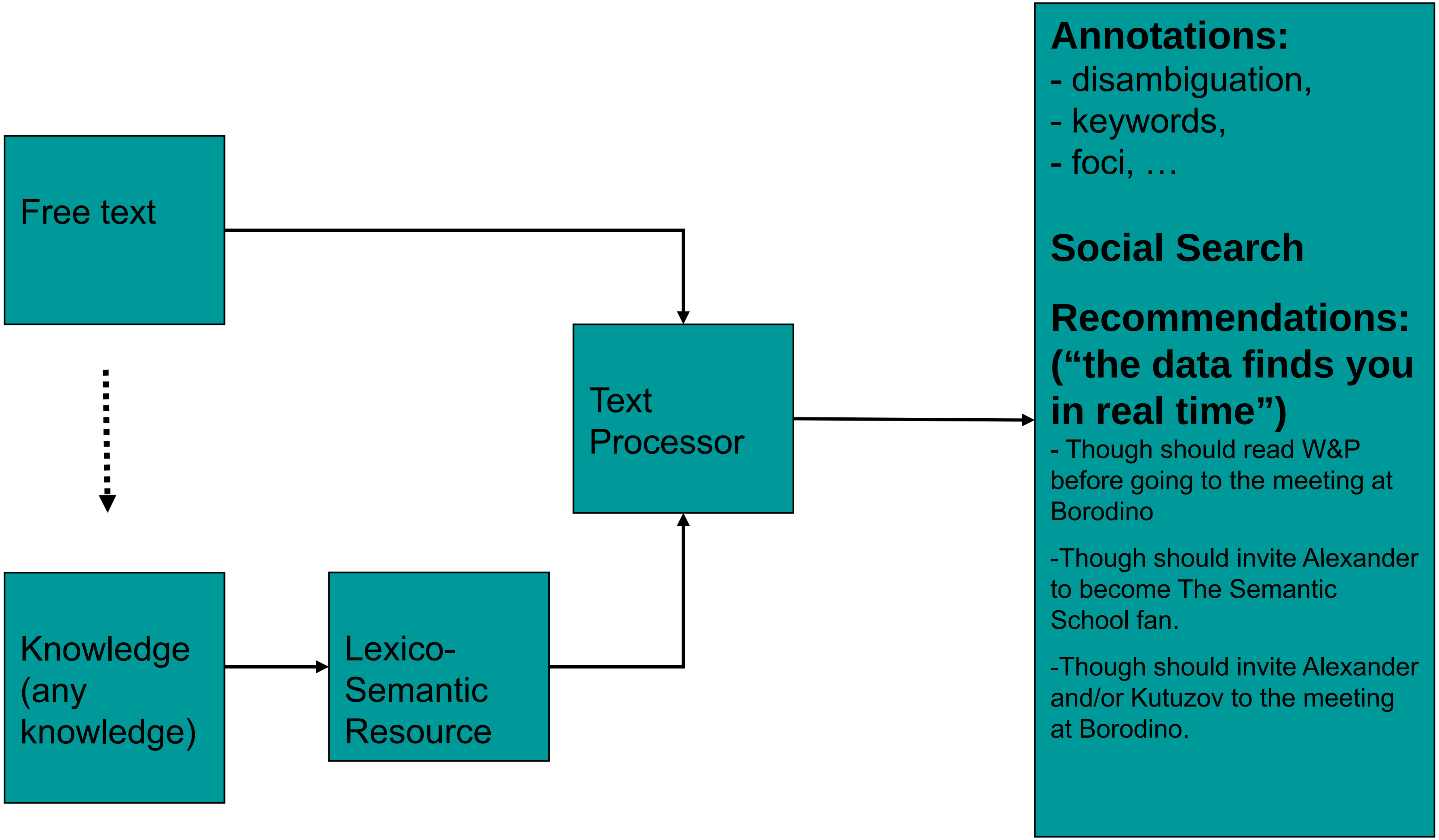
© 2010 Alexander Troussov

If weak knowledge is a good knowledge?


MY IDEA IS THAT EVERYONE SHOULD BE REQUIRED TO USE SMALL FONTS. THAT WAY WE'LL SAVE DISK SPACE.
AND I'VE NOTICED THAT MANY PEOPLE USE ENTIRE COLONS IN SITUATIONS WHERE A SEMICOLON WOULD DO JUST FINE.
S.Adams
© 1996 United Feature Syndicate, Inc. (NYC)

- COLONS ~ ONTOLOGIES
- SEMICOLONS ~ SOCIAL CONTEXT

- Suitable capital fonts and colons are hard to find, while small fonts and semicolons are now in abundance and DO JUST FINE if properly used.



Free text

Knowledge (any knowledge)

Lexico-Semantic Resource

Text Processor

Annotations:
- disambiguation,
- keywords,
- foci, …

Social Search

Recommendations: (“the data finds you in real time”)

- Though should read W&P before going to the meeting at Borodino

-Though should invite Alexander to become The Semantic School fan.

-Though should invite Alexander and/or Kutuzov to the meeting at Borodino.

© 2010 Alexander Troussov

New types of text annotations are needed

- Semantics
- Social
- Activities management
  - – This is about evocation, “getting things done”, action management, etc

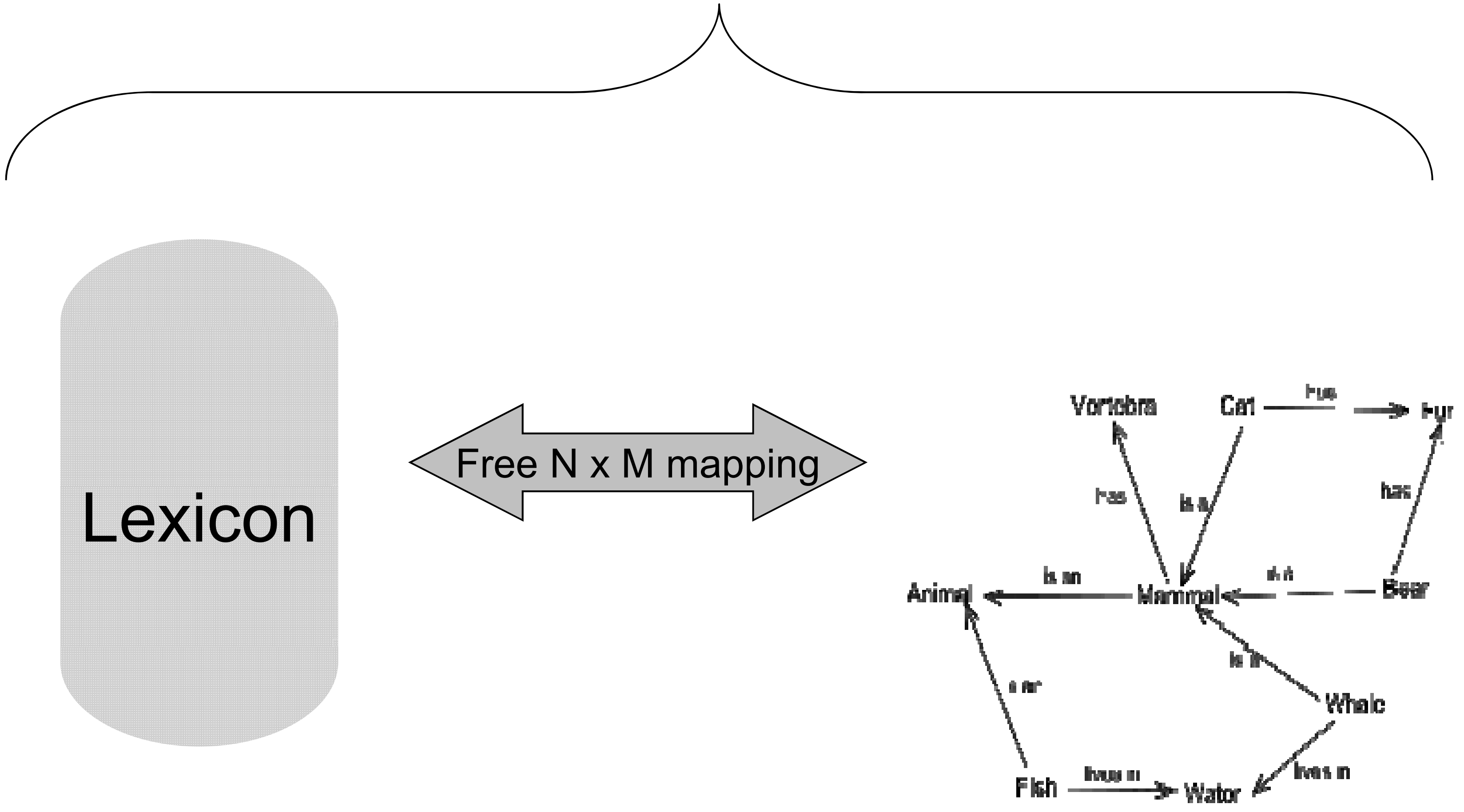
© 2010 Alexander Troussov

Knowledge and Lexico-Semantic Resources

© 2010 Alexander Troussov

Layered organisation

Layered organisation of lexico-semantic knowledge
for automatic text processing



Lexicon

Free N x M mapping

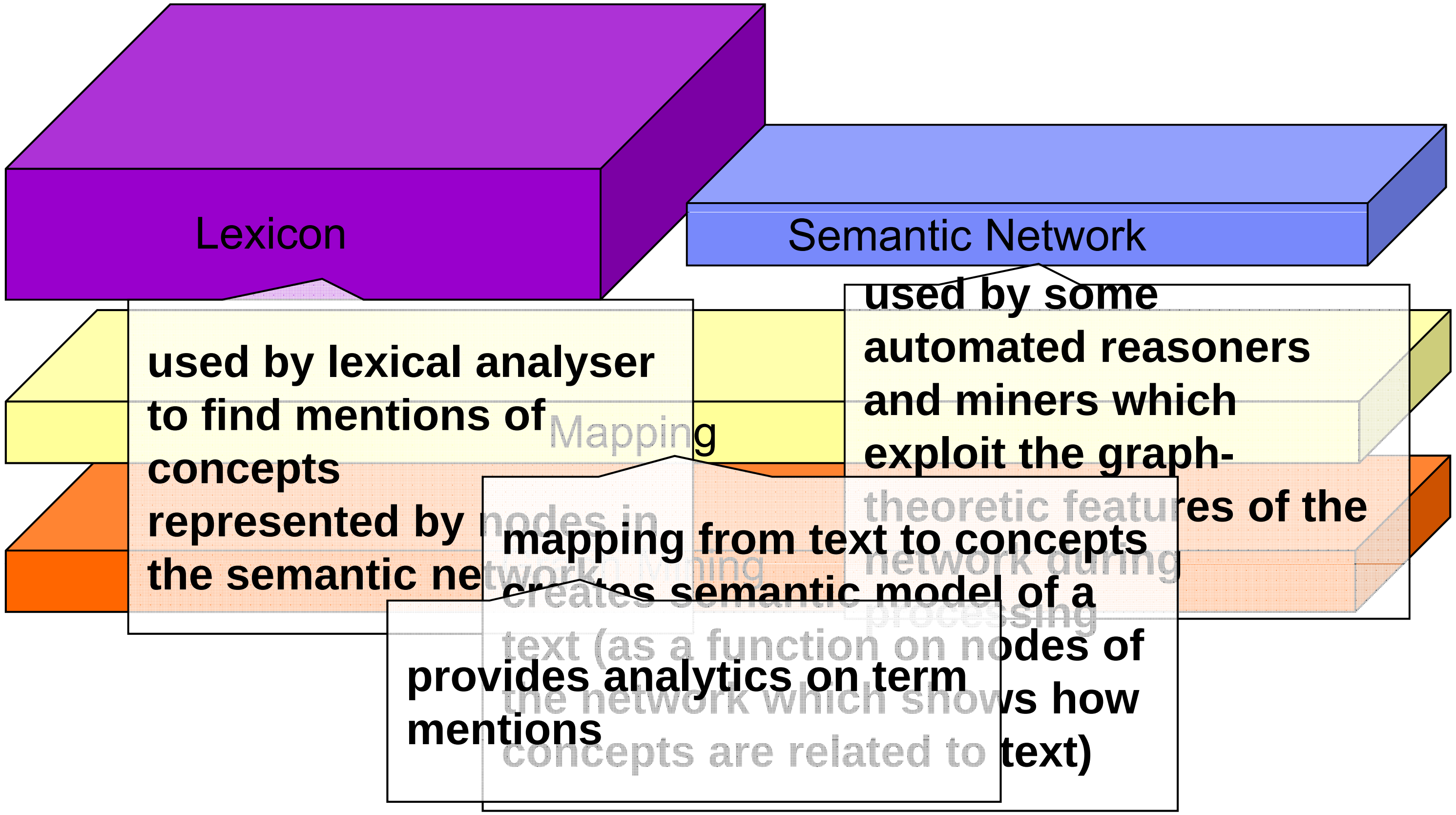
© 2010 Alexander Troussov

Layered organisation

- Key principles of the organization of lexico-semantic knowledge into a lexically enriched ontology for automatic text processing
implemented in IBM Galaxy:
  - Use of two layers: lexical entries and concepts
  - No distinctions between conceptual layer and instances layer
    - which are merged into semantic network
  - Labels != lexical entries
  - Lexical entries – any names of concepts, terms, MWU, …
  - Free *N* x *M* mappings between lexical expressions and concepts

© 2010 Alexander Troussov

… and processing resources



used by lexical analyser to find mentions of concepts represented by nodes in the semantic network

used by some automated reasoners and miners which exploit the graph-theoretic features of the network during processing

mapping from text to concepts creates semantic model of a text (as a function on nodes of the network which shows how concepts are related to text)

provides analytics on term mentions

© 2010 Alexander Troussov

Text processing

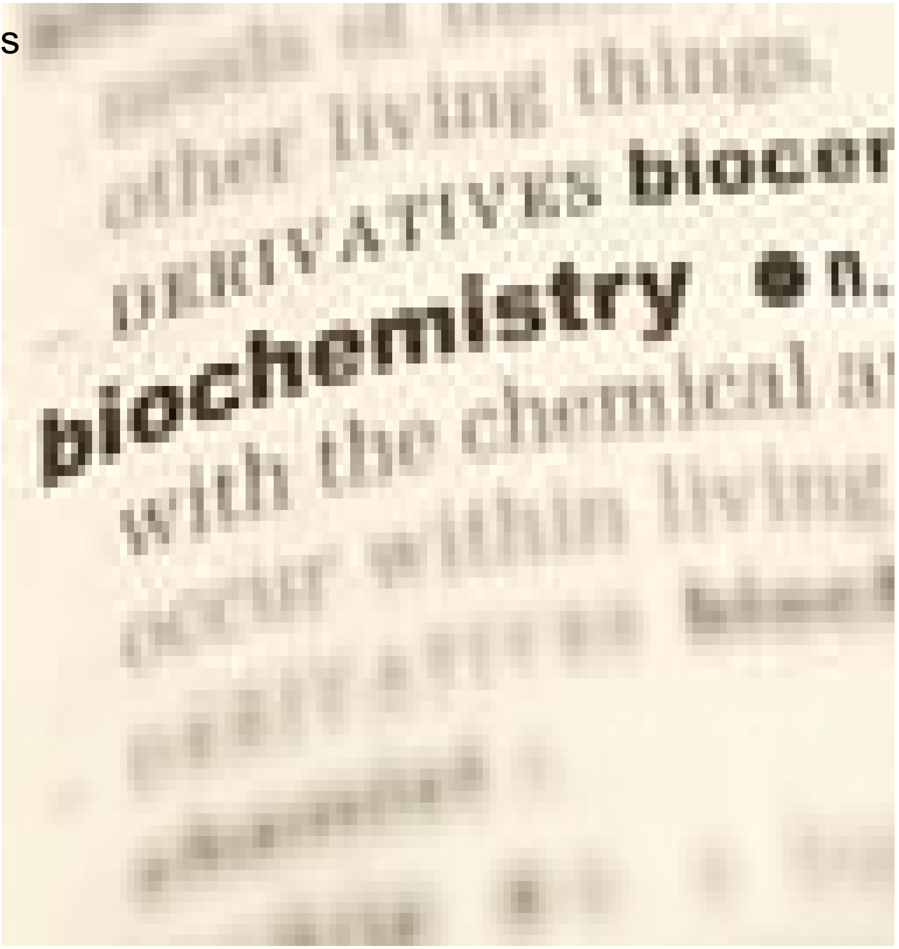
© 2010 Alexander Troussov

Knowledge-based text processing

- IBM library Galaxy is an example of robust analytics on term mentions

- Galaxy as a product for semantic text analysis:
  - Reads the text
  - Memorises all the concepts
  - Uses networks of words
    to analyse which concepts
    sit well together

- Text which mentions
  Mulhuddart, Lansdowne, Clontarf
  is probably about
  Dublin / Ireland / Europe



© 2010 Alexander Troussov

Semantic Function Space Models for Texts

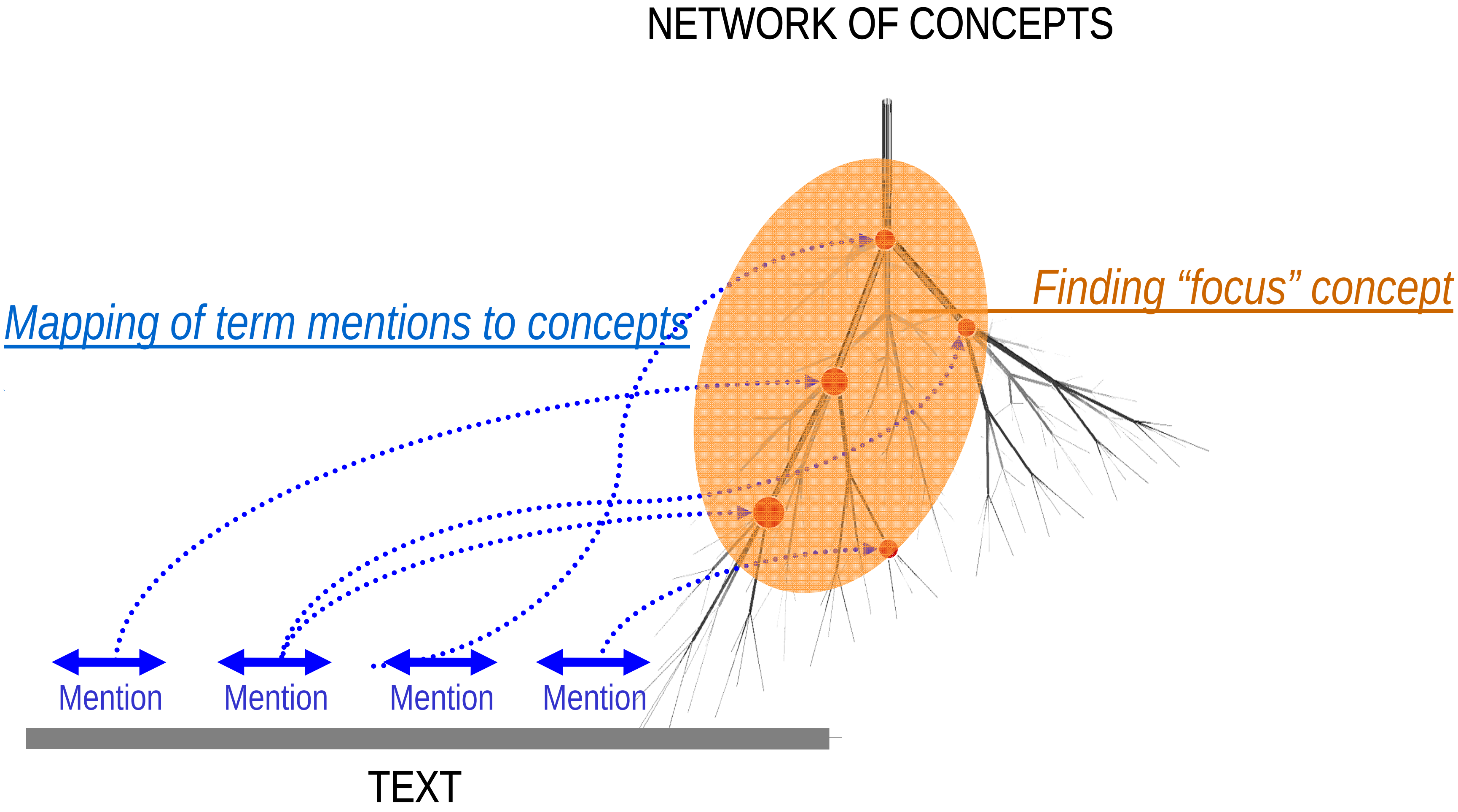
© 2010 Alexander Troussov

Semantic Function Space Models for Representing Text Documents

- TRADITIONAL: Vector Space Model (VSM)
    - Traditional Vector Space Model of Information Retrieval
- NOVEL: Semantic Function Space Model
    - Model we introduce which covers Vector Space Models and is somewhat similar to it
    - However, VSM is an algebraic model,
    - while Function Space Model can be studied by the methods of function analysis: make function more smooth, find local maximums, etc. involving graphmining

© 2010 Alexander Troussov

How it works: modelling and mining



NETWORK OF CONCEPTS

Mapping of term mentions to concepts

Finding “focus” concept

Mention

Mention

Mention

Mention

TEXT

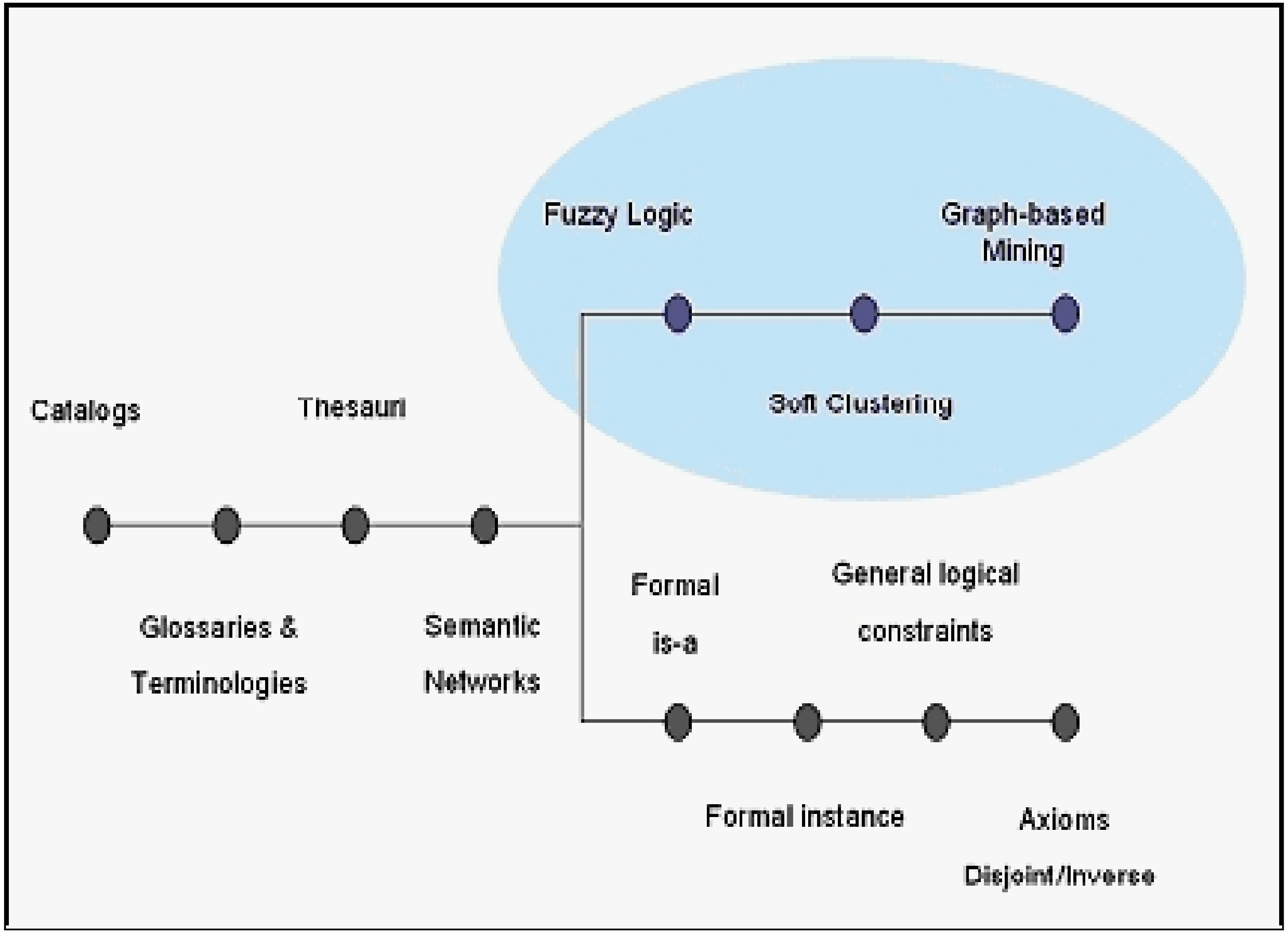
© 2010 Alexander Troussov

Semantic Function Space Model

- We model a text by mapping term mentions onto the knowledge network
- Cohesive coherent text is different from random list of terms. Smoothing the function helps to understand topicality,
  finding local maximums helps us to see the foci of the text
- Later we will talk about diffusion processes which allow to “smooth” the model

© 2010 Alexander Troussov

Graph-based approach to use “knowledge”



© 2010 Alexander Troussov

“Strong” and “Weak” Knowledge

© 2010 Alexander Troussov

What Knowledge should we use for text analytics?

- It depends …
- and this question should be properly addressed only in the view of the traditional dichotomy between top-down and bottom-up approaches in knowledge management

© 2010 Alexander Troussov

Good knowledge

- Should be relevant to the domain of the discourse
- Should know the realities of the domain
- Should be up-to-date

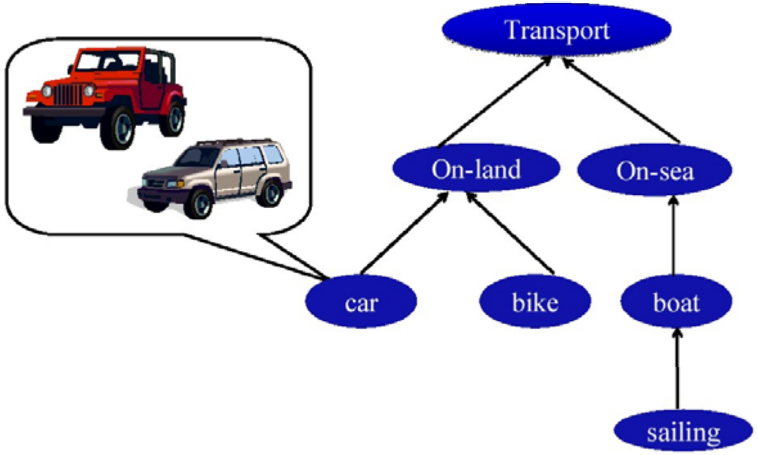
© 2010 Alexander Troussov

Where we can find “the right” knowledge?

- Where we can find “the right” knowledge for text processing?
  - Right for the domain
    - Knowledge about Symantec Client Firewall is not suitable to process W&P
  - Right for computers
- In the Social Knowledge

© 2010 Alexander Troussov

NLU as inferencing



The concept of a car is relevant to a text.
Car IS-A “on-land travel” (?)
Therefore “on-land travel” is somewhat relevant to the text, …

© 2010 Alexander Troussov

NLU as an inferencing

The notion of relevancy (text – concept “car”) with some ad hoc measure is introduced (explicitly or implicitly)

The concept of a car is relevant to the processed text

inferencing is used to propagate this relevancy measure to other concepts
“car” IS A “on-land travel” ? Therefore “on-land travel” is somewhat relevant, … (ok)

Problems
“car” IS A “on-land travel”, “bike” IS A “on-land travel”
Therefore “bike” is relevant (?)
“Napoleon” IS A “person”, “Newton” IS A “person”
Therefore “Newton” is relevant (???)

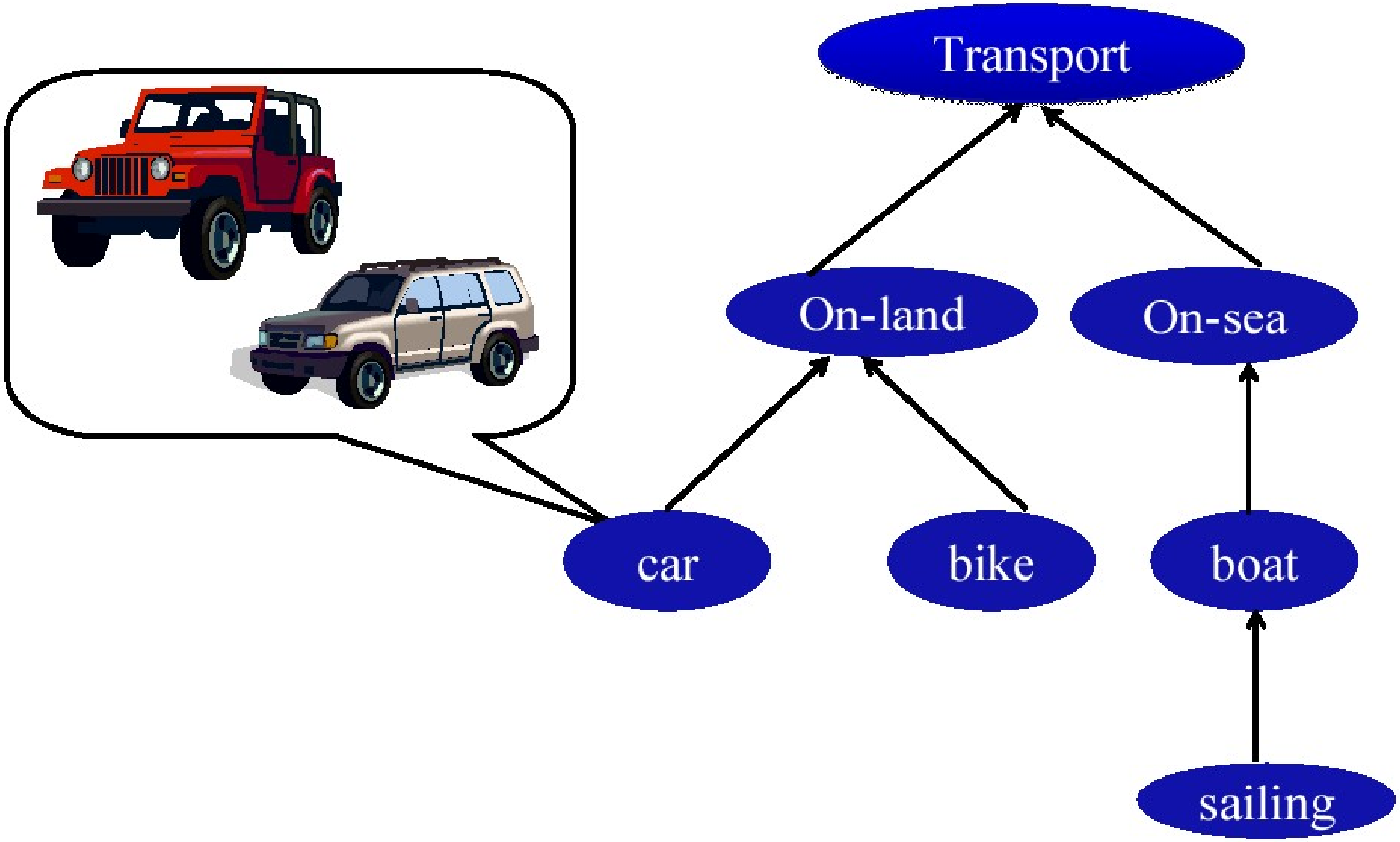
When the knowledge is “weak” -problems with inferencing become severe

Solutions:

fuzzy logic
look for inspiration from other domains:
including the use of diffusion methods to propagate
trust, knowledge, information, deceases, …

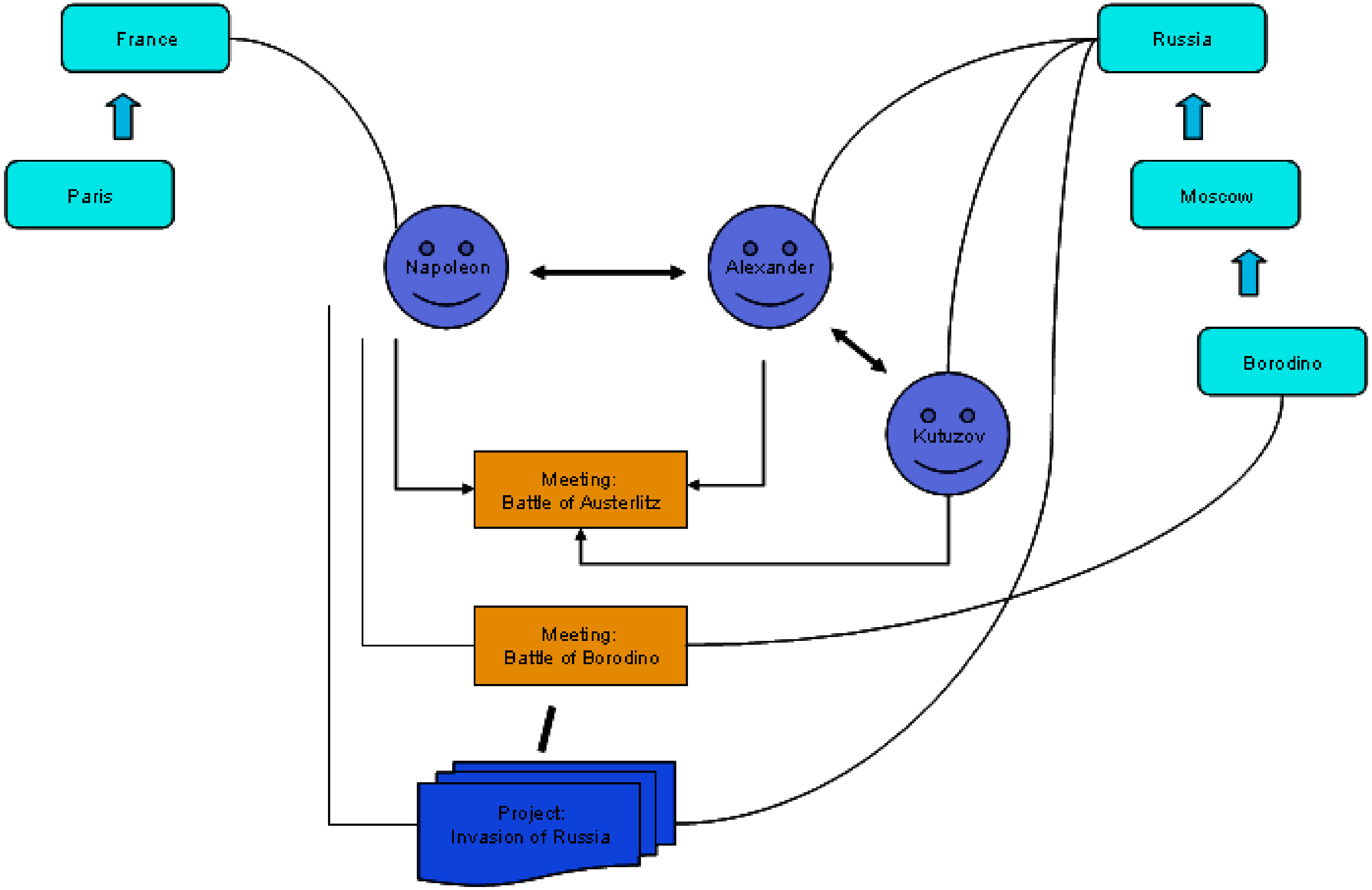
© 2010 Alexander Troussov

NLU as inferencing



© 2010 Alexander Troussov

NLU as inferencing



© 2010 Alexander Troussov

Diffusion like methods

- Diffusion –
  - the spread of social institutions (and myths and skills) from one society to another
  - (physics) the process of diffusing; the intermingling of molecules in gases and liquids as a result of random thermal agitation
  - the act of dispersing or diffusing something ("The diffusion of knowledge")
- Diffusion methods are actively used for trust, risk, web-page importance propagation.
- Following this approach:
  - "car" – relevancy 1.0
- we propagate the relevancy measure
  - "on-line transportation" – relevancy 0.5 (or 0.9, or 0.1)
  - Etc

© 2010 Alexander Troussov

Diffusion like methods (cont.)

- Diffusion-like methods are increasingly applied to social networks, hyperlink structures on the Web, electric grids etc.
  - functioning of many networks in nature is defined mainly through elementary interactions between primitive elements.
- At the same time, many of the measures in network analysis have very different interpretations in networks of different kinds.
  - Interpretations of these measures should take into account demonstrate how to make these methods aware of dimensions of networks where people are involved, including social, semantics, and activity management dimensions.
  - In the following, we will also talk about networks in general, but it should be clear from the text that many of the measures in network analysis can only be strictly interpreted in the context of social networks or have very different interpretations in networks of other kinds.

© 2010 Alexander Troussov

Diffusion-like methods

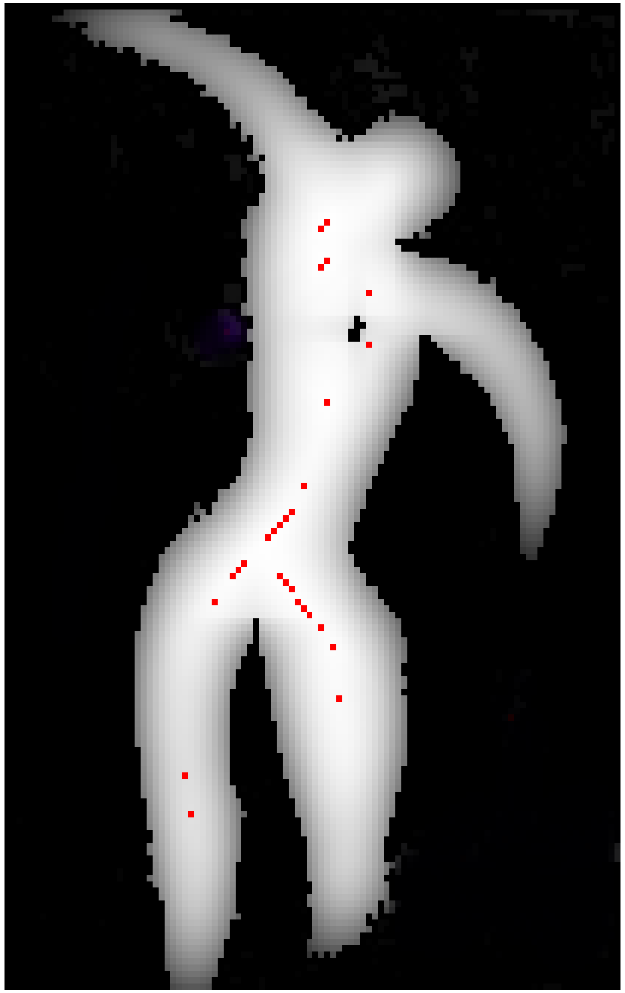
- Spreading activation is one of the diffusion-like methods. The algorithm is inspired by the phenomena observed in the nervous systems of living organisms
- In physical analogy we replace the notion of activation by the notion degree of illumination and spread of activation by the notion of light propagation
- Each node of interest within the graph emits an amount of "light" which propagates around the graph along its links
- "Light" from multiple sources combines eventually leading to a point which is illuminated to a greater degree
  - Physical analogy allows to see that the most illuminated nodes are not necessarily those nodes which were originally chosen as light emitters, but rather the overlapping areas
- This is from known to unknown, from seen to unseen
  "Je ne cherche pas, je trouve"
  Pablo Picasso

© 2010 Alexander Troussov

Are capable to analyse massive networks

- Detect structure of networks at different time slices
    - With complex topology (not necessarily “grids” as in image processing)
- Applications:
    - trend analysis,
      risk assessment
        - The component “left leg” became more prominent
        - The component “Head” become less prominent

Numerical simulation on the scanned image with 6000 pixels done by IBM Galaxy tool



© 2010 Alexander Troussov

Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

RuSSIR

Russian Summer School
in Information Retrieval

Applications of network flow graph-based methods for mining and using network models of social context

© 2010 Alexander Troussov

Previous art in use of network models of weak knowledge

- Troussov et al. "Social Context as Machine-Processable Knowledge" MITACS FP-Nets Workshop on Social Networks, August 2010, Vancouver, Canada.

  – We examined previous art in use of network models of weak knowledge (see the list on the next slide)
  – We found (demonstrated) that graph-based methods used there were used mainly for fuzzy inferencing and soft clustering.
  – and we created a set of “atomic” network flow operations
  – Which allow to generalize the approach taken in previous art

© 2010 Alexander Troussov

Summary and Conclusions (Cont.)

- The same methods actually perform well with simplified multidimensional network models of instantiations of ontologies without the need for encoding the rules how to use proper conceptualisation
- We revised previous art in use of network models of weak knowledge, and we described the algorithms and the architecture of the hybrid recommender system in the activity centric environment Nepomuk-Simple (EU 6th Framework Project NEPOMUK):
- The applications constituting previous art were monolithic software applications. In this paper we present a novel computational paradigm which breaks these applications into “atomic” components, where the computational methods for propagation are separated as distinct “atomic” network flow engines. This approach provides a unified view of previous applications. From the software engineering prospective the advantages of such an approach includes easy software maintenance, reuse and optimization of network flow engines, and the guide for new applications.
  - We created a set of “atomic” network objects
  - And the set of network flow-based operations with such objects
  - And described efficient scalable implementations of such operations.
- Performance – subsecond for networks with several hundreds K nodes

© 2010 Alexander Troussov

Representing knowledge

- There are a number of options:
  - As *objects*, using the well-accepted techniques of object-oriented analysis and design to capture a model
  - As *clauses* going back to the early days of AI and Lisp
  - As *XML*, using the industry-standard structured mark-up language
  - As *graphs*, making use of the things we know about graph theory
  - As some combination of these
- We are looking for
  - Extensibility
  - Easy of merge heterogeneous information
  - Ease of use
- And our choice is - GRAPHS
  - Knowledge is represented by a multidimentional network which is modeled by a graph
  - And the Nepomuk-Simple will give us examples of extensibility/Easy of merge heterogeneous information/Ease of use
-

Source: Simon Dobson, Declan O'Sullivan

© 2010 Alexander Troussov

Graphs

- We can use the nodes of a graph for facts, concepts, people, organisations, etc and the arcs as binary relationships between them
  - Arcs are typically called predicates or relationships in this view
  - The set of arcs intersecting a node tells us the information we know about the fact or entity

Source: Simon Dobson, Declan O'Sullivan

© 2010 Alexander Troussov

Nepomuk

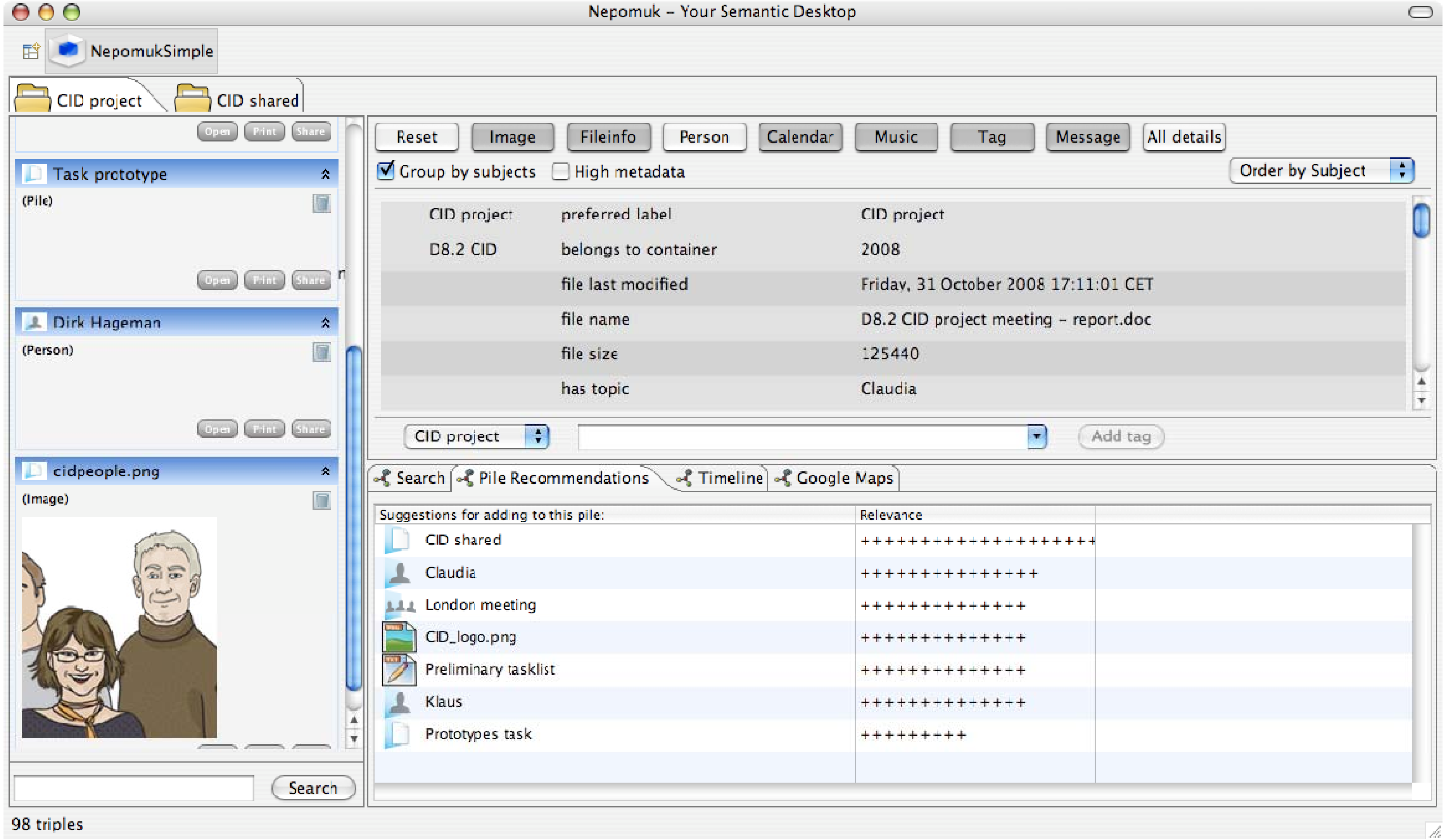
- NEPOMUK (Networked Environment for Personalized, Ontology-based Management of Unified Knowledge) is an open-source software specification that is concerned with the development of a social semantic desktop that enriches and interconnects data from different desktop applications using semantic metadata stored as RDF.
- Initially, it was developed in the EU 6th framework integrated project Nepomuk (2006-2008) - 17 million euros, of which 11.5 million was funded by the European Union
- Partners
  - German Research Center for Artificial Intelligence DFKI GmbH Germany
  - International Business Machines (IBM) Ireland
  - SAP AG Germany
  - Hewlett-Packard Galway Ltd Ireland
  - Thales SA TRT France
  - PRC Group The Management House S.A Greece
  - Mandriva Edge-IT France
  - Cognium Systems SA France
  - National University of Ireland, Galway Ireland
  - Ecole Politechnique Fédérale de Lausanne Switzerland
  - Forschungszentrum Informatik an der Universität Karlsruhe Germany
  - L3S Research Center Germany
  - Institute of Communication and Computer Systems of the National Technical University of Athens, Greece
  - Kungliga Tekniska Högskolan Sweden
  - Università de la Svizzera Italiana Switzerland
  - Irion Management Consulting GmbH Germany

© 2010 Alexander Troussov

Nepomuk hybrid recommender

- We present the architecture of the hybrid recommender system in the activity centric environment Nepomuk-Simple (EU 6th Framework Project NEPOMUK).
- “Real” desktops usually have piles of things on them where the users (consciously or unconsciously) grouped together items which are related to each other or to a task. The so called “Pile” UI, used in the Nepomuk-Simple imitates this type of data and metadata organization which helps to avoid premature categorization and reduces the retention of useless documents.
- Metadata describing the user data are stored in the Nepomuk personal information management ontology (PIMO). Proper recommendations, such as recommendation of additional items to add to the pile, apparently should be based on the PIMO, on the textual content of the items in the pile. Although methods of natural language processing for information retrieval could be useful, the most important type of textual processing are those which allows to related concepts in PIMO to the processed texts. Since PIMO changes over the time, this type of natural language processing can’t be performed as preprocessing of all textual context related to the user. Hybrid recommendation needs on-the fly textual processing with the ability to aggregate the current instantiation of PIMO with the results of textual processing.
- Representing and modeling this ontology as a multidimensional network allows to augment the ontology on the fly by new information, such as the “semantic” content of the textual information in user documents. Recommendations in the Nepomuk-Simple are computed on the fly by graph-based methods performing in the unified multidimensional network of concepts from the personal information management ontology augmented with concepts extracted from the documents pertaining to the activity in question. In this paper, we classify Nepomuk-Simple recommendations into two major types. The first type of recommendations is recommendation of the additional items to the pile, when the user is working on an activity. The second type of recommendations arises, for instance, when the user is browsing Web; the Nepomuk-Simple can recommend that current resource might be relevant to one or more activities performed by the user. In both cases there is a need to operate with Clouds (fuzzy sets of PIMO nodes): Clouds describe topicality of documents in terms of PIMO, the pile itself is a Cloud.

Pile UI



Task prototype (Pile)

Dirk Hageman (Person)

cidpeople.png (Image)

☑ Group by subjects □ High metadata

Order by Subject

CID project	preferred label	CID project
D8.2 CID	belongs to container	2008
	file last modified	Friday, 31 October 2008 17:11:01 CET
	file name	D8.2 CID project meeting – report.doc
	file size	125440
	has topic	Claudia

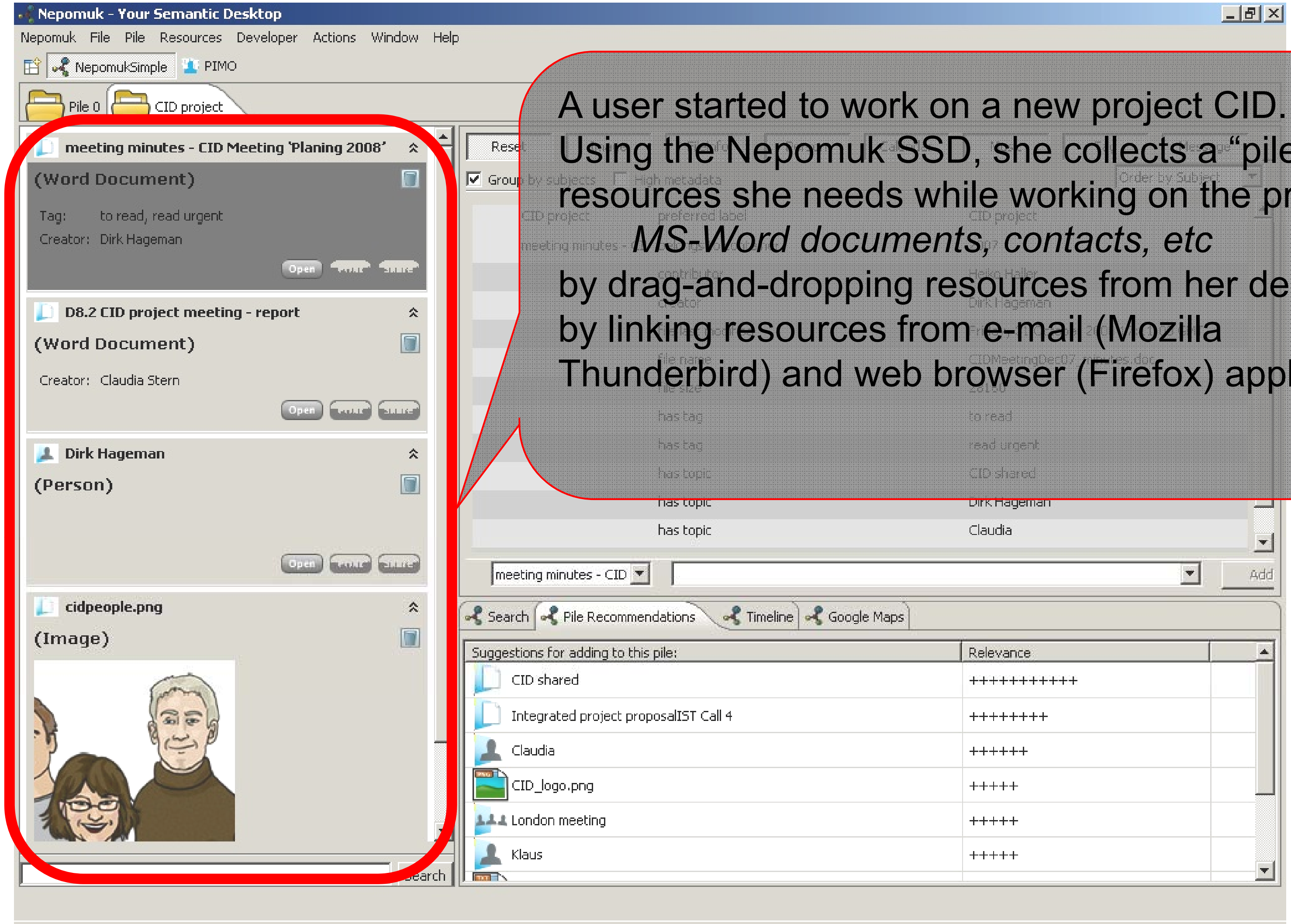
CID project

Suggestions for adding to this pile:	Relevance
CID shared	+++++++++++++++++++
Claudia	++++++++++++++
London meeting	+++++++++++++
CID_logo.png	+++++++++++++
Preliminary tasklist	+++++++++++++
Klaus	+++++++++++++
Prototypes task	+++++++++

98 triples

© 2010 Alexander Troussov

Nepomuk use case: activity management



meeting minutes - CID Meeting 'Planing 2008'
(Word Document)
Tag: to read, read urgent
Creator: Dirk Hageman

D8.2 CID project meeting - report
(Word Document)
Creator: Claudia Stern

Dirk Hageman
(Person)

cidpeople.png
(Image)

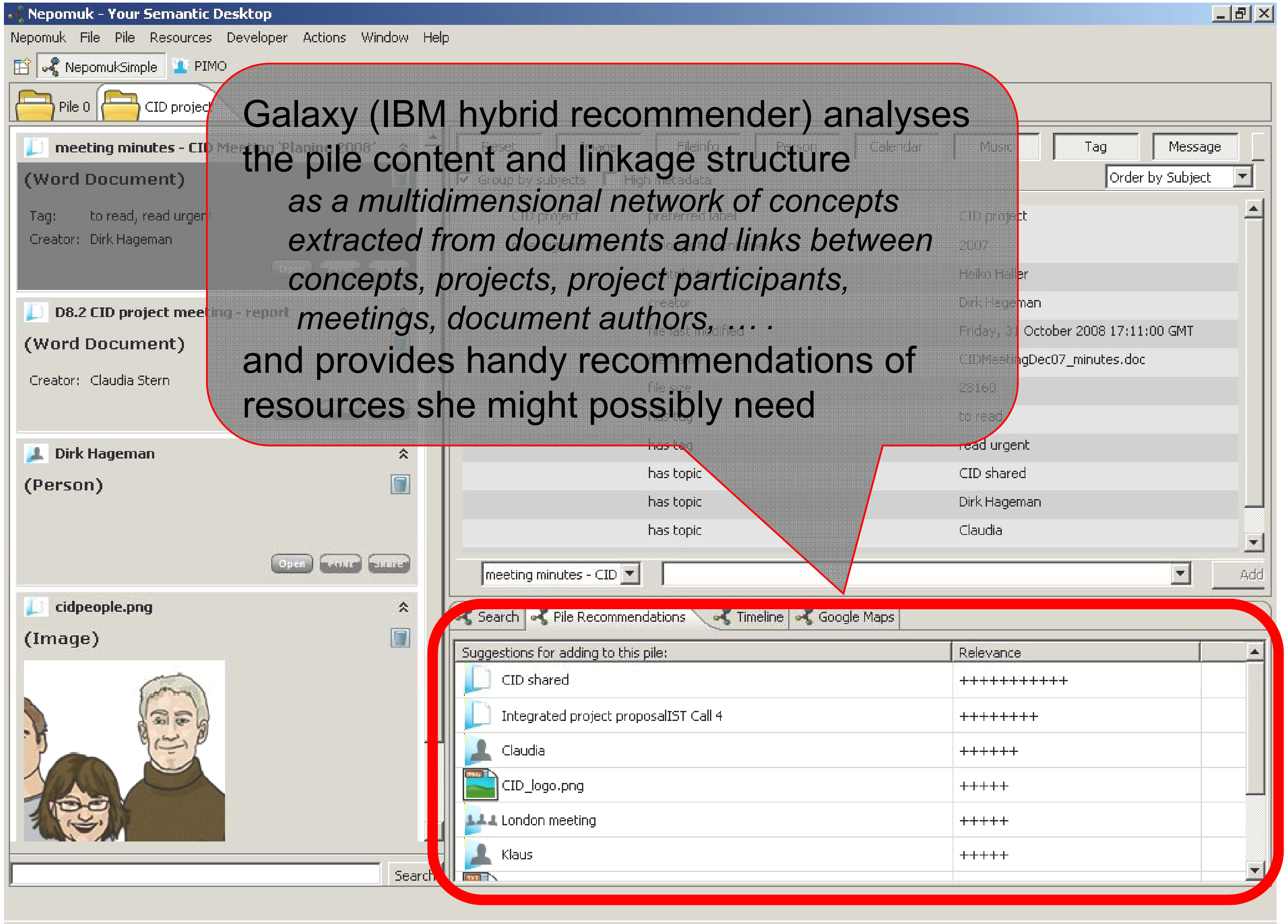
has tag	to read
has tag	read urgent
has topic	CID shared
has topic	Dirk Hageman
has topic	Claudia

Suggestions for adding to this pile:	Relevance
CID shared	+++++++++++
Integrated project proposalIST Call 4	++++++++
Claudia	++++++
CID_logo.png	+++++
London meeting	+++++
Klaus	+++++

A user started to work on a new project CID.
Using the Nepomuk SSD, she collects a "pile" of resources she needs while working on the project:
MS-Word documents, contacts, etc
by drag-and-dropping resources from her desktop,
by linking resources from e-mail (Mozilla Thunderbird) and web browser (Firefox) applications.

© 2010 Alexander Troussov

Nepomuk use case: activity management



Galaxy (IBM hybrid recommender) analyses the pile content and linkage structure

as a multidimensional network of concepts extracted from documents and links between concepts, projects, project participants, meetings, document authors, … .

and provides handy recommendations of resources she might possibly need

meeting minutes - CID Meeting 'Planing 2008'

(Word Document)

Tag: to read, read urgent

Creator: Dirk Hageman

D8.2 CID project meeting - report

(Word Document)

Creator: Claudia Stern

Dirk Hageman

(Person)

cidpeople.png

(Image)

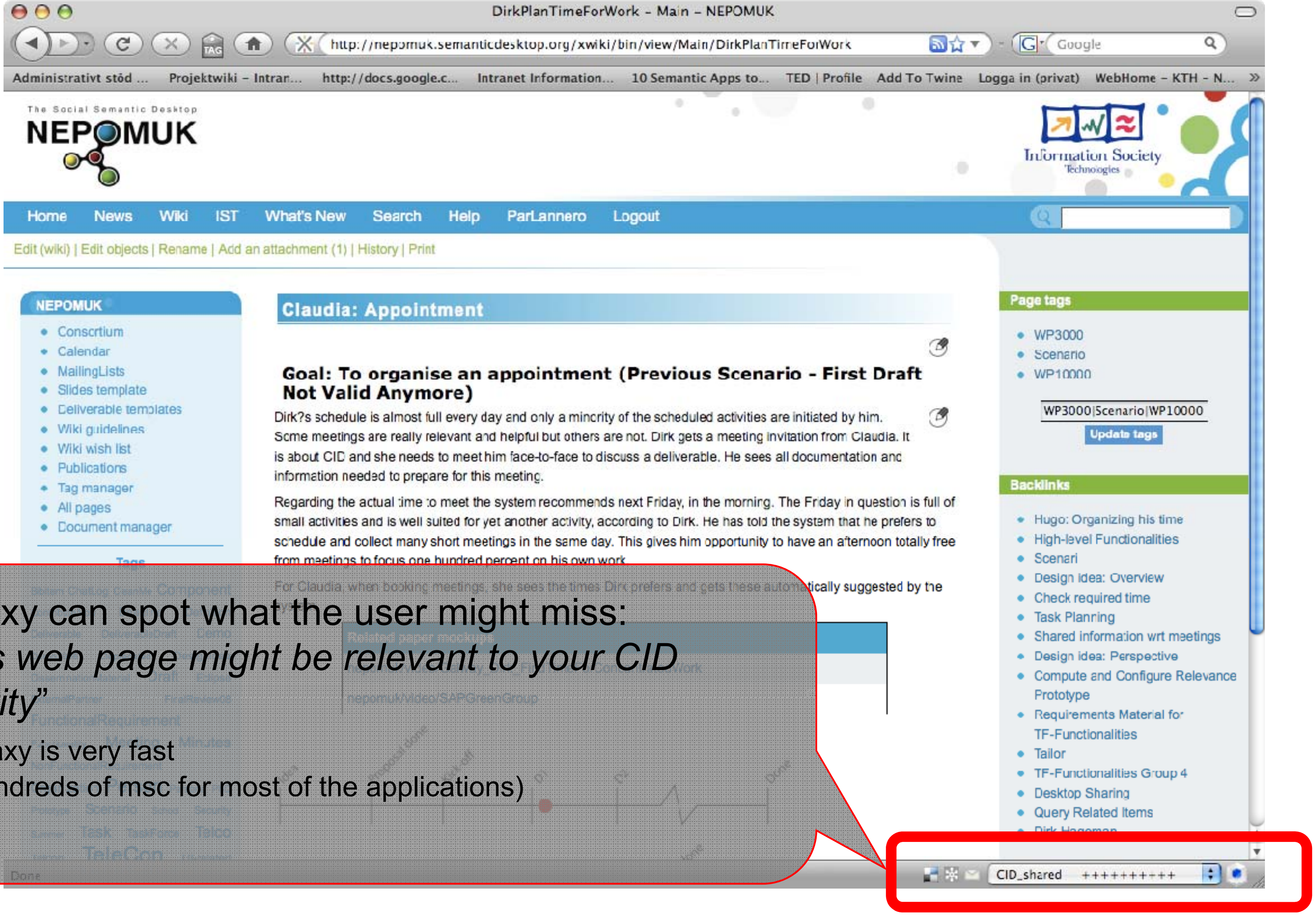
preferred label	CID project
	2007
	Heiko Haller
creator	Dirk Hageman
file last modified	Friday, 31 October 2008 17:11:00 GMT
	CIDMeetingDec07_minutes.doc
file size	28160
has tag	to read
has tag	read urgent
has topic	CID shared
has topic	Dirk Hageman
has topic	Claudia

meeting minutes - CID

Suggestions for adding to this pile:	Relevance
CID shared	+++++++++++
Integrated project proposalIST Call 4	++++++++
Claudia	++++++
CID_logo.png	+++++
London meeting	+++++
Klaus	+++++

© 2010 Alexander Troussov

Nepomuk use case: activity management



DirkPlanTimeForWork - Main - NEPOMUK

http://nepomuk.semanticdesktop.org/xwiki/bin/view/Main/DirkPlanTimeForWork

Google
Administrativt stöd ... Projektwiki - Intran... http://docs.google.c... Intranet Information... 10 Semantic Apps to... TED | Profile Add To Twine Logga in (privat) WebHome - KTH - N...
The Social Semantic Desktop
NEPOMUK
Information Society Technologies
Home News Wiki IST What's New Search Help ParLannero Logout
Edit (wiki) | Edit objects | Rename | Add an attachment (1) | History | Print
NEPOMUK
Consortium
Calendar
MailingLists
Slides template
Deliverable templates
Wiki guidelines
Wiki wish list
Publications
Tag manager
All pages
Document manager
Tags

Claudia: Appointment

Goal: To organise an appointment (Previous Scenario - First Draft Not Valid Anymore)

Dirk?s schedule is almost full every day and only a minority of the scheduled activities are initiated by him. Some meetings are really relevant and helpful but others are not. Dirk gets a meeting invitation from Claudia. It is about CID and she needs to meet him face-to-face to discuss a deliverable. He sees all documentation and information needed to prepare for this meeting.

Regarding the actual time to meet the system recommends next Friday, in the morning. The Friday in question is full of small activities and is well suited for yet another activity, according to Dirk. He has told the system that he prefers to schedule and collect many short meetings in the same day. This gives him opportunity to have an afternoon totally free from meetings to focus one hundred percent on his own work.

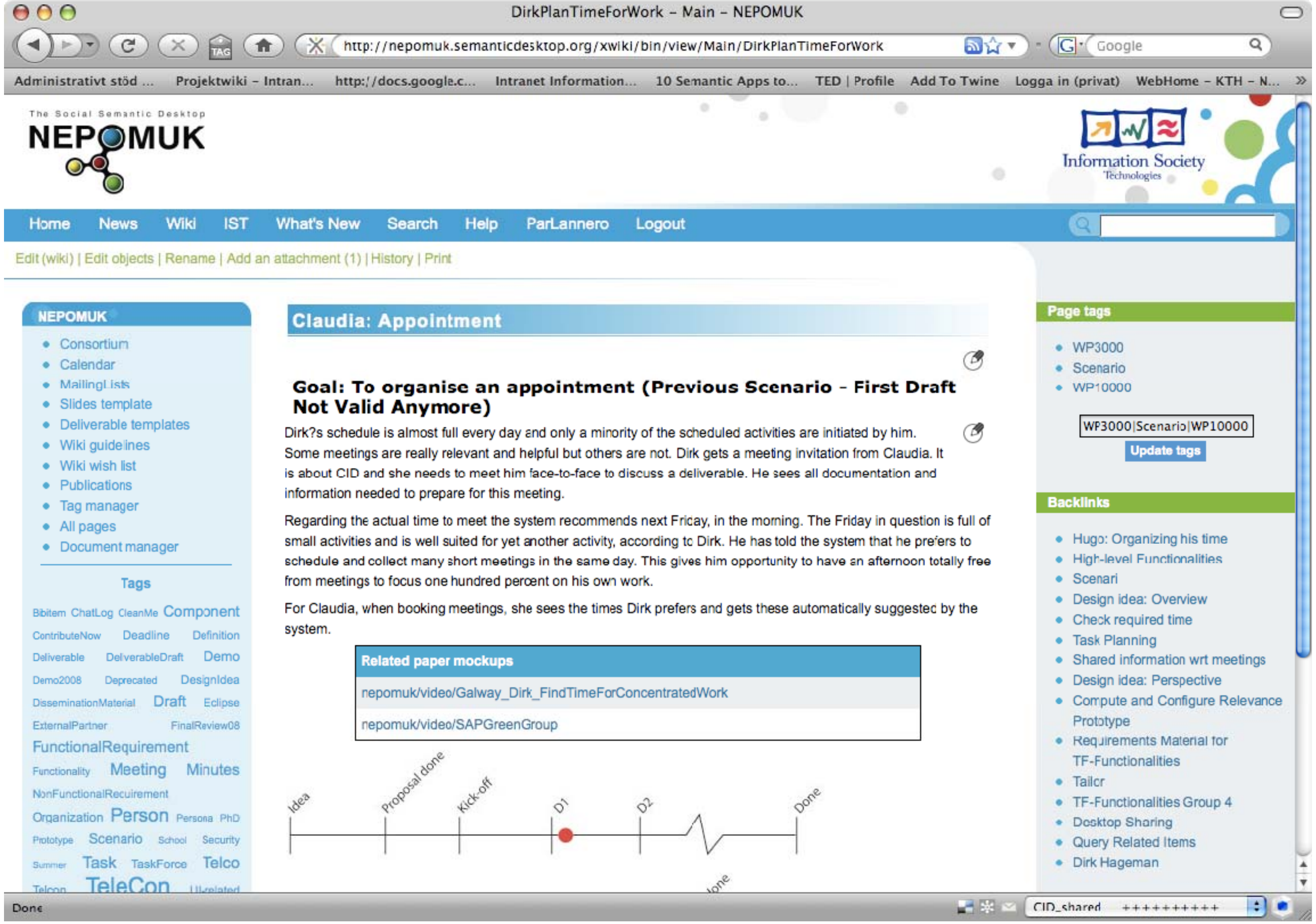
For Claudia, when booking meetings, she sees the times Dirk prefers and gets these automatically suggested by the

Galaxy can spot what the user might miss:
“*This web page might be relevant to your CID activity*”

- Galaxy is very fast
  (hundreds of msc for most of the applications)

© 2010 Alexander Troussov

Pile recommendation for a webpage


DirkPlanTimeForWork - Main - NEPOMUK
http://nepomuk.semanticdesktop.org/xwiki/bin/view/Main/DirkPlanTimeForWork
Google
Administrativt stöd ...
Projektwiki - Intran...
http://docs.google.c...
Intranet Information...
10 Semantic Apps to...
TED | Profile
Add To Twine
Logga in (privat)
WebHome - KTH - N...
The Social Semantic Desktop
NEPOMUK
Information Society Technologies
Home
News
Wiki
IST
What's New
Search
Help
ParLannero
Logout
Edit (wiki) | Edit objects | Rename | Add an attachment (1) | History | Print
NEPOMUK
Consortium
Calendar
MailingLists
Slides template
Deliverable templates
Wiki guidelines
Wiki wish list
Publications
Tag manager
All pages
Document manager
Tags
Bibitem ChatLog CleanMe Component ContributeNow Deadline Definition Deliverable DeliverableDraft Demo Demo2008 Deprecated DesignIdea DisseminationMaterial Draft Eclipse ExternalPartner FinalReview08 FunctionalRequirement Functionality Meeting Minutes NonFunctionalRequirement Organization Person Persona PhD Prototype Scenario School Security Summer Task TaskForce Telco Telcon TeleCon Unrelated
Page tags
WP3000
Scenario
WP10000
WP3000|Scenario|WP10000
Update tags
Backlinks
Hugo: Organizing his time
High-level Functionalities
Scenari
Design idea: Overview
Check required time
Task Planning
Shared information wrt meetings
Design idea: Perspective
Compute and Configure Relevance Prototype
Requirements Material for TF-Functionalities
Tailor
TF-Functionalities Group 4
Desktop Sharing
Query Related Items
Dirk Hageman
Done
CID_shared ++++++++++

Claudia: Appointment

Goal: To organise an appointment (Previous Scenario - First Draft Not Valid Anymore)

Dirk?s schedule is almost full every day and only a minority of the scheduled activities are initiated by him. Some meetings are really relevant and helpful but others are not. Dirk gets a meeting invitation from Claudia. It is about CID and she needs to meet him face-to-face to discuss a deliverable. He sees all documentation and information needed to prepare for this meeting.

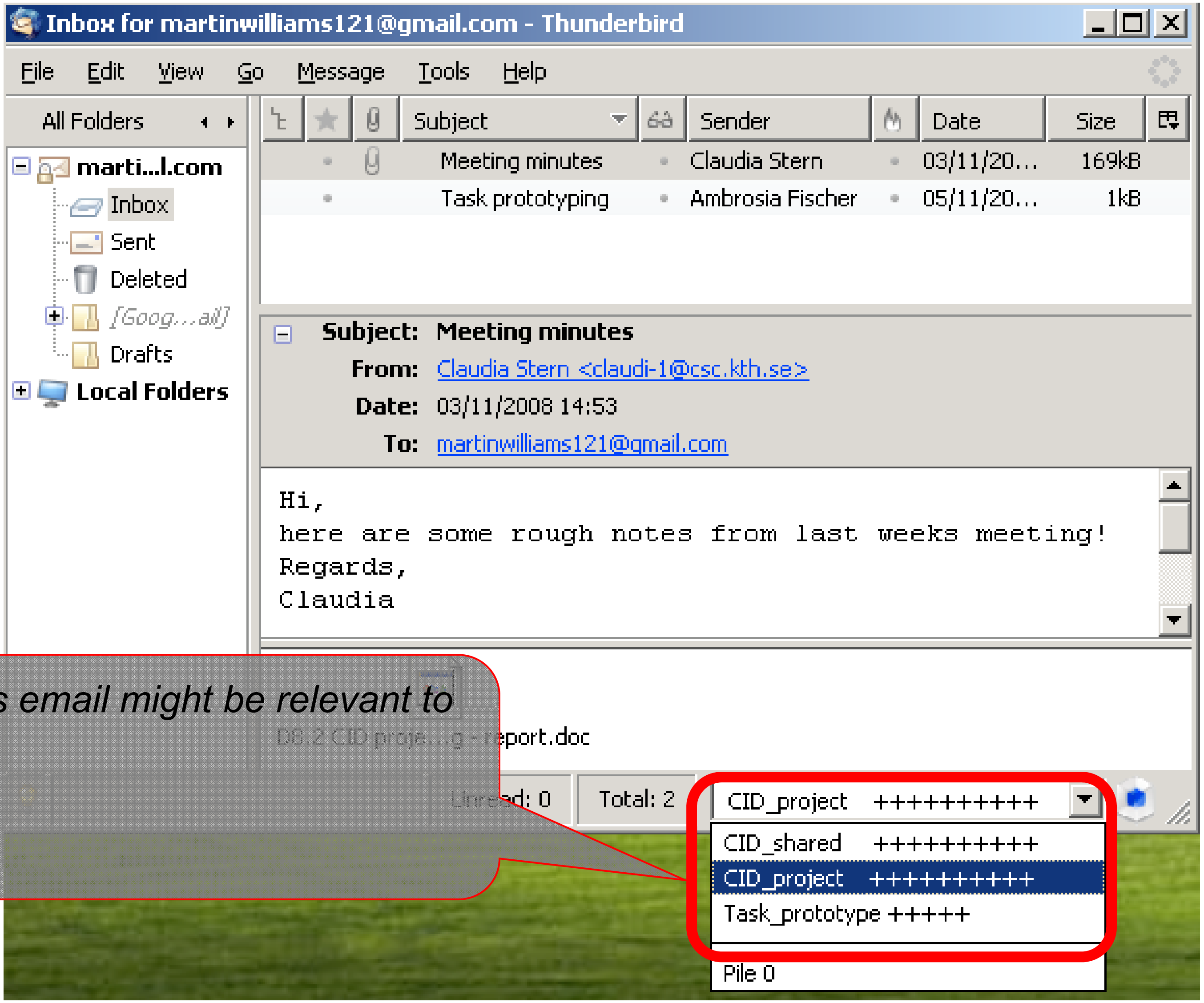
Regarding the actual time to meet the system recommends next Friday, in the morning. The Friday in question is full of small activities and is well suited for yet another activity, according to Dirk. He has told the system that he prefers to schedule and collect many short meetings in the same day. This gives him opportunity to have an afternoon totally free from meetings to focus one hundred percent on his own work.

For Claudia, when booking meetings, she sees the times Dirk prefers and gets these automatically suggested by the system.

Related paper mockups
nepomuk/video/Galway_Dirk_FindTimeForConcentratedWork
nepomuk/video/SAPGreenGroup

© 2010 Alexander Troussov

Nepomuk use case: activity management



Inbox for martinwilliams121@gmail.com - Thunderbird
File Edit View Go Message Tools Help
All Folders
marti...l.com
Inbox
Sent
Deleted
[Goog...ail]
Drafts
Local Folders

Subject	Sender	Date	Size
Meeting minutes	Claudia Stern	03/11/20...	169kB
Task prototyping	Ambrosia Fischer	05/11/20...	1kB

Subject: Meeting minutes
From: Claudia Stern <claudi-1@csc.kth.se>
Date: 03/11/2008 14:53
To: martinwilliams121@gmail.com

```
Hi,
here are some rough notes from last weeks meeting!
Regards,
Claudia
```

D8.2 CID proje...g - report.doc

"This email might be relevant to …"

© 2010 Alexander Troussov

Network flow methods

- Social context modeled as a multidimensional network can be used as an efficient machine processable knowledge representation for various tasks. The application of this method includes such traditional areas of knowledge usage as knowledge based text understanding, and the recently emerged area of recommender systems.
- We analyse several applications and show that computational methods used in these applications are based on the network flow process, “that focuses on the outcomes for nodes in a network where something is flowing from node to node across the edges” (Borgatti and Everett, M. 2006]
- We interpret this “something” as a relevancy measure; for instance, the initial seed input value which shows nodes of interest in the network. Propagating the relevancy measure through outgoing links allows us to compute the relevancy measure for other network nodes and dynamically rank these nodes according to the relevancy measures.
- The same paradigm could be used to address the centrality measurements in social network analysis. Centralisation of the network can be achieved when we assume that all the nodes are equally important, and iteratively recompute the relevancy measure based on the connections between nodes. In addition to “global” centralisation, “local” centralisation could be performed if the initial seed values represent the nodes of interest.

© 2010 Alexander Troussov

Network flow as relevancy propagation / redistribution

- Our definition of a network flow:
  - A discrete process when on each iteration the value of a relevancy measure at a node is recomputed based on the connections between nodes.
- Dual interpretation:
  - Relevancy measure could be interpreted as a membership function which defines the fuzzy set of nodes. In this interpretation the network flow provides transformation of the fuzzy set into the sequence of fuzzy sets (expanding, shrinking,…). When the operation is “expanding” the process is usually called spreading activation.
  - Alternatively, the process is iterative redistribution of the relevancy measure and frequently is the process of computation numerical approximation to the solution of a partial differential equation
-

© 2010 Alexander Troussov

Network objects – operands of network flow

- Object (Network object) – is a node or a (fuzzy) set of nodes on the network . Fuzzy sets are characterized by a membership function *M* which shows the degree of belongings of an element to the set.
- We also use term cloud where we want to emphasize the fact that the membership function is non-negative real-valued function, not Boolean valued.

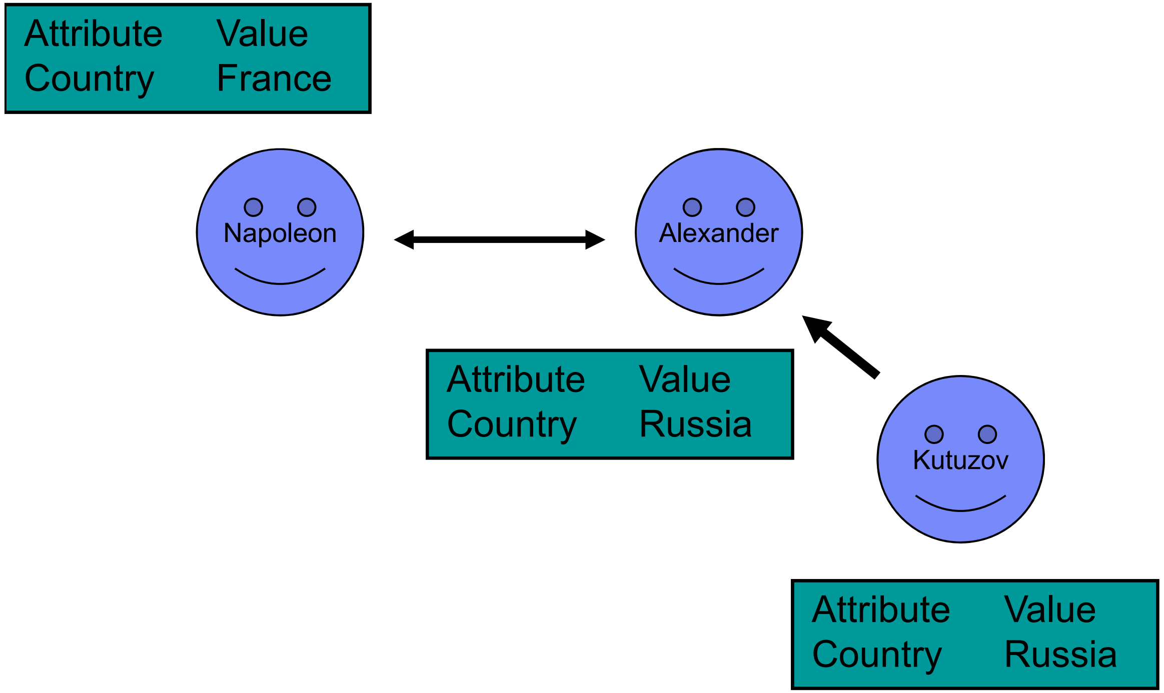
© 2010 Alexander Troussov

Clouds

- Clouds per se are not part of the network, although they could be encoded into the topology of the network

© 2010 Alexander Troussov

Attribute–value pairs



Attribute	Value
Country	France

Napoleon

Alexander

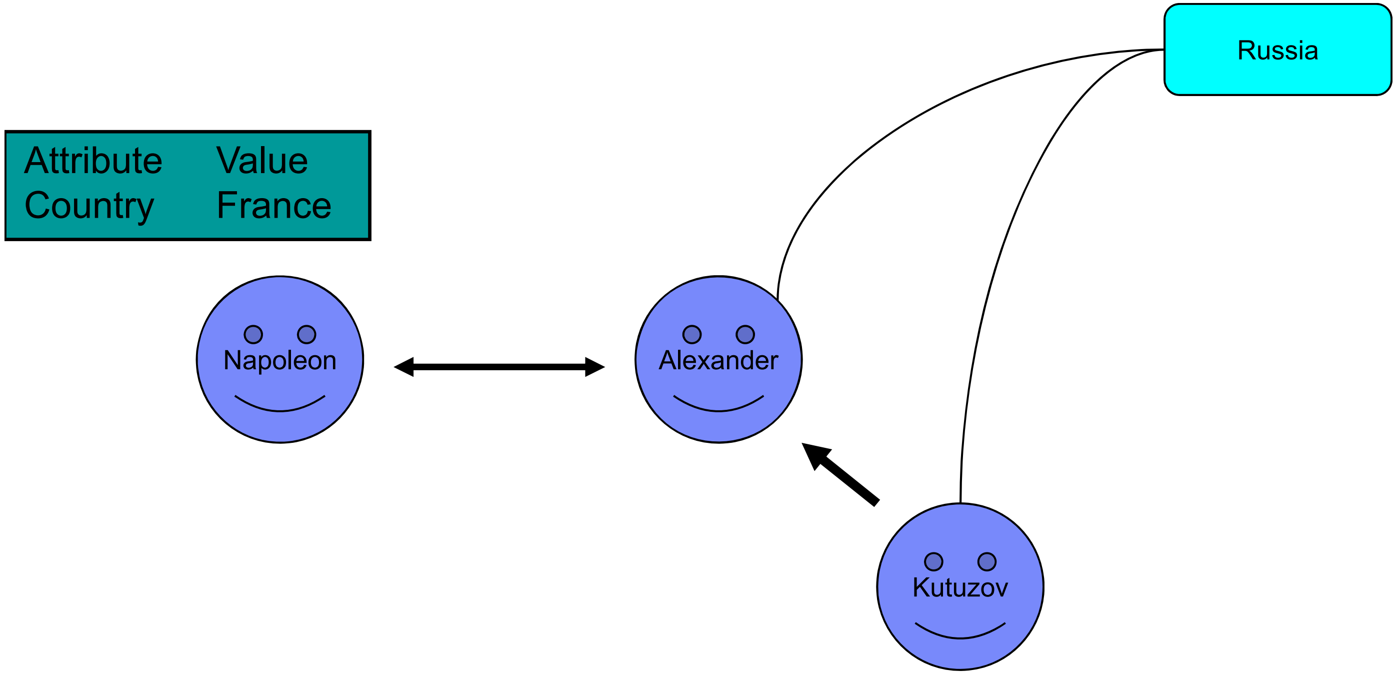
Attribute	Value
Country	Russia

Kutuzov

Attribute	Value
Country	Russia

© 2010 Alexander Troussov

Attribute–value pairs → Topology



© 2010 Alexander Troussov

Clouds (cont.)

- Clouds frequently are task specific and created dynamically:
  - – Text could be modeled as cloud of concepts from a static semantic network
  - – Activities (as in the Nepomuk-Simple)

© 2010 Alexander Troussov

Notations

- *Object* (Network object) – is a node or a (fuzzy) set of nodes on the network. Fuzzy sets are characterized by a membership function *M* which shows the degree of belongings of an element to the set.
- *M* – the membership function for fuzzy sets which is a non-negative real-valued function.
- *Activation* – the membership function when it is not interpreted in the fuzzy sets paradigm. We use the activation (the activation of nodes, or objects) as an abstract relevancy measure.
- *Cloud* (cloud object) – we use the term cloud where we want to emphasize the fact that the membership function is non-negative real-valued function, not Boolean valued. As usual, we assume that a node *e* belongs to the fuzzy set *C,* or in mathematical notations $e \in C$ - if M(e) ≥ 0.
- |…| - cardinality of sets. In case of clouds we define $|C|= \Sigma\ M(e)$ for all nodes *e* such that $e \in C$.
- *Query* – an object used as a seed for local ranking (defined below)

© 2010 Alexander Troussov

Operations with one argument

- *Expansion* – is a unary operation which transform a cloud into another cloud: *Expansion*: $C1 \rightarrow C2$. If $C1$ and $C2$ are crisp sets, we assume that $C1$ is a proper subset of $C2$: $C1 \subset C2$. If general, we assume that this operation increases the number of the nodes with non-zeroed membership function values, doesn't change significantly the values of the membership functions on the nodes in C1, and that $|C1| \leq |C2|$.
  - *Useful to compare "sparse" clouds*
- *Smoothing* – is formally the same as expansion, however the interpretation of this operation can't be done in the framework of fuzzy sets, instead, it roots in the operations with functions in calculus. We assume that smoothing makes the difference between the values of the function M() on neighbor nodes smaller.
  - *a blend of fuzzy inferencing and soft clustering useful for text processing)*
- *Local ranking* - is formally the same as expansion. The purpose of this operation is to get the value of the activation which shows the proximity, or relevance, of objects to a query.
- *Shrinking* – is a unary operation which transform a cloud into another cloud: *Shrinking*: $C2 \rightarrow C1$. If $C1$ and $C2$ are crisp sets, we assume that $C1$ is a proper subset of $C2$, i.e. $C1 \subset C2$. If general, we assume that this operation decreases the number of the nodes with non-zeroed membership function values, doesn't change significantly the values of the membership functions on the nodes in $C1$, and that $|C2| \leq |C1|$. Shrinking is a kind of inverse operation to expansion, although we don't necessarily assume that for any pair of such operations $C1 \equiv Shrinking(Expansion(C1)$ for each object $C1$.

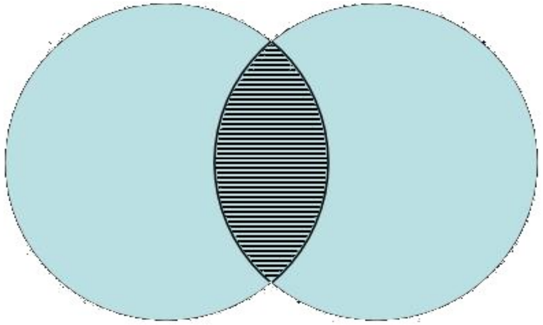
© 2010 Alexander Troussov

Operations with two and more arguments – Similarity, Search

- Similarity (Dissimilarity) of Sets of Nodes and Search for Similar Sets
  - Discussion on the distinction between similarity and proximity of network nodes is outside of the scope of this paper. In this Section we present empirical approach to computation of similarity based on a network flow process. Similarity of network nodes, or more generally similarity of two network objects (like clouds which are fuzzy sets of network nodes) could be described in terms of their ability to affect various parts of the network (like in viral marketing applications. In other words, similarity of two sets A0 and B0 should be defined as similarity of two fuzzy sets A=*Expanding*(A0) and B=*Expanding*(B0), where the operation *Expanding* is done by network flow methods compatible with the targeted applications. For instance, if the target applications is in the area of “viral marketing”, than we expect that the *Expanding* is done by network flow methods which model “viral marketing”.
- In section 4.1 we provide additional arguments to justify out approach to similarity and introduce the similarity of two fuzzy sets on a network. In Section 4.2 we describe efficient and scalable implementation of search for similar sets.

© 2010 Alexander Troussov

Similarity metrics in Set Theory



- Venn Diagram

- Set theory”: The measure of similarity - how much in common two sets have is measured in terms of an "exact" match. The similarity value is a number in the range 0 to 1, 0 – no common elements, 1 – the sets are equal.
- Using fuzzy logic operations we define similarity as:

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

© 2010 Alexander Troussov

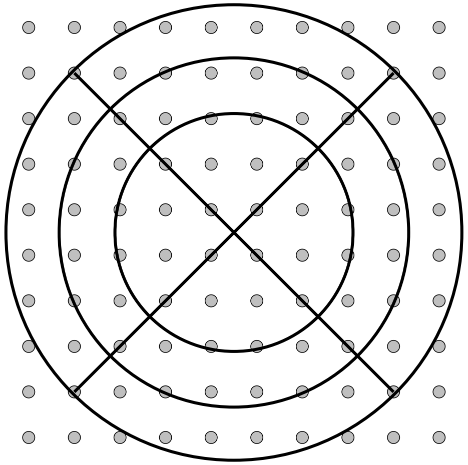
Similarities/Dissimilarities in Set Theory

- Similarity of users in techno-social systems is used for so called community based recommendations.
- Techno-social systems are different:
    - In contrast to the behavior of users in classic recommender systems, the users of bookmarking systems are opportunistic - users randomly bookmark items rather than trying to rate as much as they can as in classic recommenders.
        - It means that the fact that two users bookmarked same 10 items is important, but the fact that each has 50 other items not bookmarked by the other is not that important. Usage of popular similarity measures (like cosine similarity, Pearson correlation and Jaccard index [6]) to measure similarity between users of collaborative tagging systems provides the results which are overly affected by dissimilarities.
- In this section we propose a novel method which allows flexibility for taking into account similarities/dissimilarities.
    - Firstly, users are modeled by fuzzy sets of related nodes on a network model. Secondly, comparison of user models is done by the use of fuzzy logic which allows us to control the importance . We define a new metric for nodes similarity which shows flexibility for taking into account similarities/dissimilarities. With some choose of logical AND and OR operations the above mentioned formula gives the results insensitive to the dissimilarities.

© 2010 Alexander Troussov

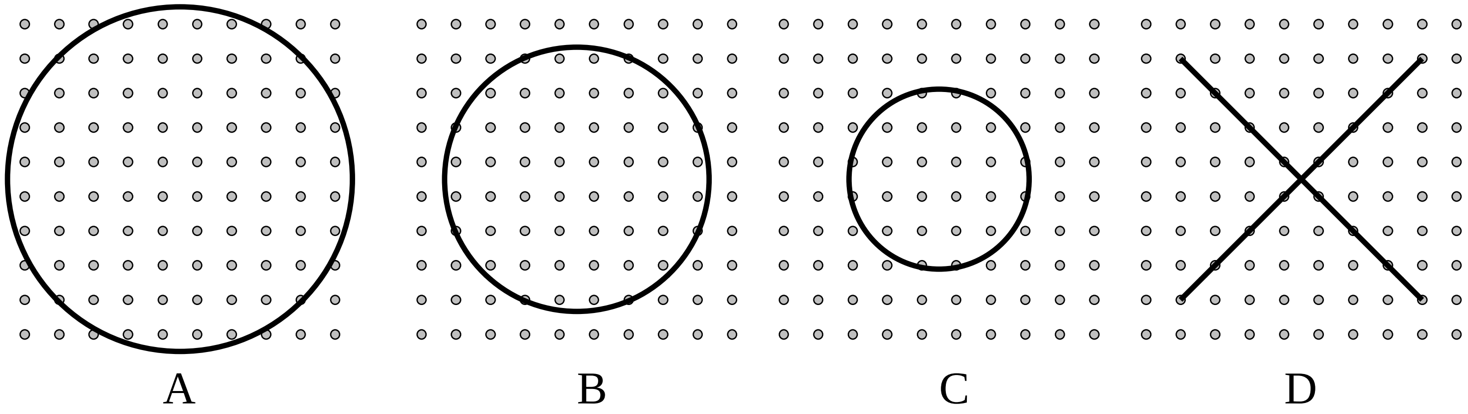
Similarity between two groups of nodes

- However, our sets are actually sets of nodes in a network of concepts. Because the links between concepts indicate the semantic proximity of concepts (including synonymy relations), these relations must be taken into account when comparing the sets. Instead of the degree of exact match, we need to use a "fuzzy" matching technique.
- To illustrate this “fuzzy” matching, let us consider a contrived geometrical example, where the network of concepts is the grid on two dimensional plane like the one on the Fig. 2.



© 2010 Alexander Troussov

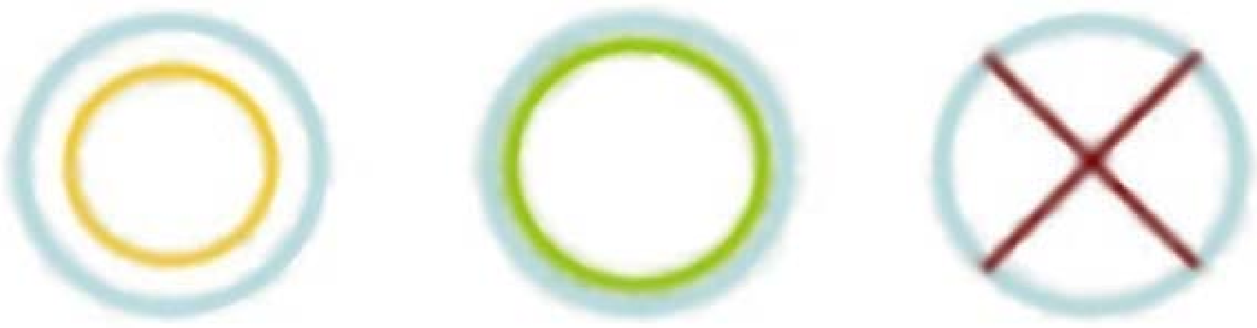
- Let us consider four sets of nodes (semantic models) with the shape depicted on the Fig 3 (with centers of symmetry place at the origin of the grid)



  - Which two sets of nodes are most similar? If our matching strategy is to look for an exact match, then the pair A and D would be most similar because they have the most nodes in common. However, intuitively, A and B are closest. How do we make a computation based on this intuition which will show us that A and B are very similar?

© 2010 Alexander Troussov

- Our approach for "fuzzy" matching is to expand all the sets by making their boundaries less well defined and more "fluffy" and as the measure of similarity between original pair we choose the exact match (i.e. overlap) of their expanded variants.



- Fig. 4. To provide "fuzzy" comparison of original sets, we perform their “fuzzyfication” first by the operation which we denoted previously as *Expanding*.
- From the Fig 4 we see that A and C still do not have common area, and hence the measure of their similarity is still zero. Sets A and B became practically indistinguishable, and their measure of similarity is close to 1, while sets A and D have common areas only in the four corners.

© 2010 Alexander Troussov

- The method we suggested to compare two “clouds” (fuzzy sets of nodes) is very close to the simplified version described above. The difference is that the nodes in the original sets are provided with membership function

© 2010 Alexander Troussov

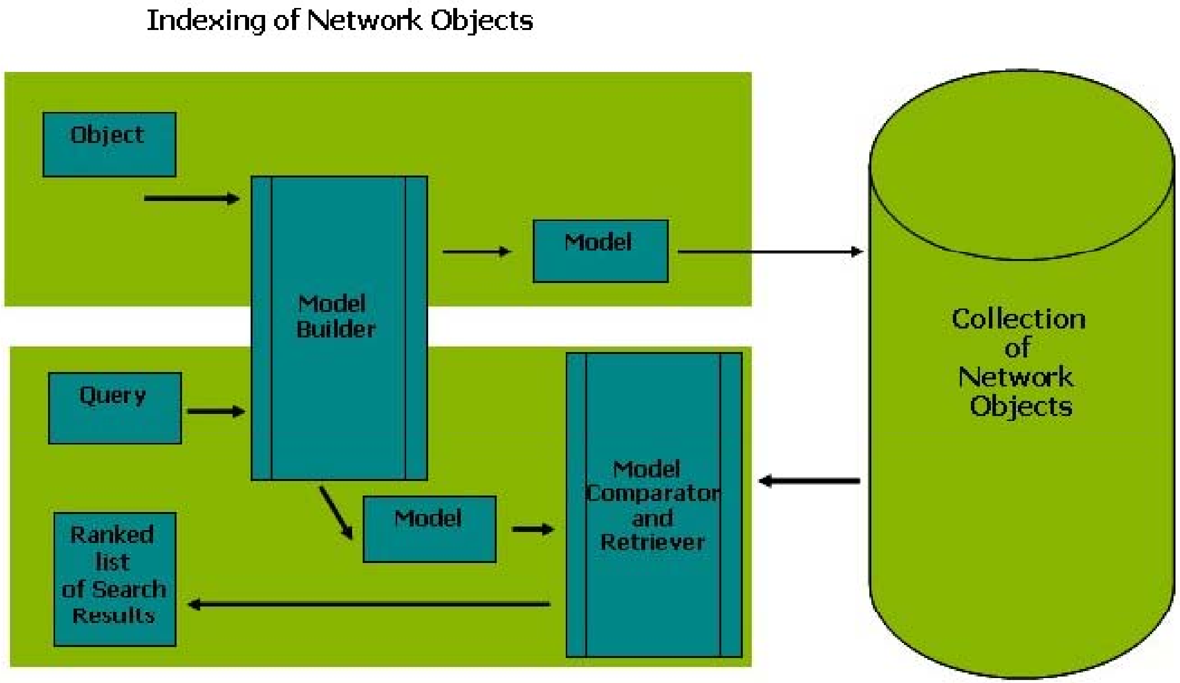
Querying Collection of N Clouds
retrieval of a similar sets in a collection - n-ary operation

- Computing of similarity of two sets above - is really fast (100msc)
- But retrieval of a similar cloud from a collection of 100 clouds will be 100 times slower if done straightforwardly. We created scalable solution Here we briefly outline that method.

© 2010 Alexander Troussov

Querying Collection of N Clouds (Cont.)

- The retrieval process takes a user query as input. A semantic model of the query is created by the model builder. This query model is then passed to the retrieval module which compares the query model to the document collection and retrieves a ranked list of documents according to the set similarity measures described above.



Indexing of Network Objects
Object
Model Builder
Model
Collection of Network Objects
Query
Model
Model Comparator and Retriever
Ranked list of Search Results

© 2010 Alexander Troussov

Outline of the algorithm

- **Step 1.** Building the model of a cloud.
- This technical step aimes to increase the storage capacity by modeling a cloud A by *Shrinking*(A); alternatively, the operation *Expanding*(A) could be used to improve recall.
- **Step 2.** This model is added to the repository ((collection) of processed network objects, which is a node-by-object matrix, each element of which is the weight of node *i* in the object *j*.
- **Step 3.** The model *Expanding*(B) of a query B can be compared against other models stored in the repository. Similarity score of document *j* and document *k* is calculated, for instance, using cosine similarity function:

$$S(d_j, d_k) = \frac{\sum_{i=1}^{N} w_{ij} \cdot w_{ik}}{\sqrt{\sum_{i=1}^{N} w_{ij}^2} \cdot \sqrt{\sum_{i=1}^{N} w_{ik}^2}}$$

- If row-wise matrix storage is used, semantic models that do not have common non-zero weight nodes with the target semantic model (and therefore guaranteed to produce zero similarity), are eliminated before a similarity score is calculated, this speeds processing.

Network-flow based Computational Systems for Mining and Use of the Social Context

- In this Section we present novel software architecture for mining and use of network models of social context based on a set of atomic software engines implementing one of the basic network flow operations described above. Arguments (operands) of these operations are network objects which we define as fuzzy sets of network nodes.
- This architecture generalizes the design of systems constituting the previous art without introducing new components which could potentially hamper performance and scalability. We show that efficient and scalable implementations for each of the atomic software engines actually exist (although as part of monolithic software applications). For instance, the paper Judge et al. 2007 describes the system which perform atomic operations on network with several hundreds nodes in 200msc on an ordinary PC. The paper Troussov et al. 2008 describes the large scale multifunctional application where various recommendations are done using hybrid methods including natural text processing. Therefore we conclude that the architecture described in this Section could be successfully used to mine and exploit the social context.

Judge, J., Sogrin, M., and Troussov, A. "Galaxy: IBM Ontological Network Miner". *Proceedings of the 1st Conference on Social Semantic Web (CSSW),* September 26-28, 2007, Leipzig, Germany.
Troussov, A., Judge, J., Sogrin, M., Bogdan, C., Lannero, P., Edlund, H., and Sundblad, Y. "Navigation Networked Data using Polycentric Fuzzy Queries and the Pile UI Metaphor". *Proceedings of the International SoNet Workshop (2008)*, pp. 5-12.

© 2010 Alexander Troussov

Network-flow based Computational Systems for Mining and Use of the Social Context (Cont.)

- Major steps involved in building the software application based on principles described in this paper are:
  - Modeling the social context (such as instantiations of techno-social systems) using multidimensional networks.
  - Task is modeled as a cloud - fuzzy set of nodes which performs the role of the *Query*
  - Task dependent enhancement of the model of the social context
    - the network is enhanced on the fly by new objects and new links between node, and augmented by new task dependent objects
  - Local ranking using *Query* as the initial seed provides the ranked list of network objects relevant to the *Query*
- The previous sections provide examples of these steps. For instance, in the Nepomuk-Simple the underlying network is enhanced on the fly by concepts extracted from the textual content of pile items.

© 2010 Alexander Troussov