Module Web Intelligence (WI)

Chapitre I : Introduction

Pr. Okba KAZAR

Professeur des universités
Directeur du Laboratoire d'INFormatique Intelligente
LINFI

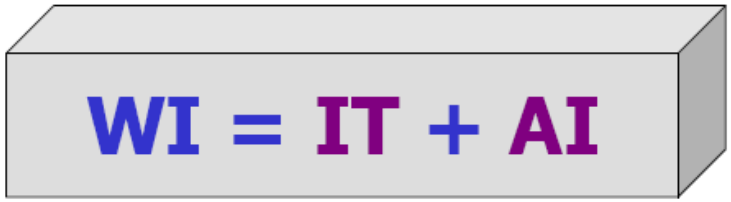
Département d'informatique
Université de biskra

Contenu

- Une brève histoire du Web Intelligence
- Motivations pour WI
- Définition et perspectives de WI
- Programme de recherche
- Principaux outils de Web Intelligence
- conclusion

C’est quoi “Web Intelligence”

- **Web Intelligence (WI) exploite l'impact fondamental et pratique que la technologie de l'information de pointe (TI) et l’innovante intelligence artificielle (IA) auront sur le Web**



  - Integration de IT avec IA
  - Applications de IA sur le Web

Une brève histoire du Web Intelligence

- 1999: initiatives de recherche Collaborative
  - Ning Zhong, Data Mining et Systemes à base de connaissances
  - Jiming Liu, agents Intelligent et système multi-agents
  - Yiyu Yao, La recherche d'information et les systèmes d'information intelligents

- Les efforts de recherche combinés avec objectif commun: créer une nouvelle sous-discipline couvrant les théories et les techniques liées à l'information Web .

Une brève histoire du Web Intelligence

- 2000: Publication d'un document de deux pages sur la position sur WI (Zhong, Liu, Yao, Ohsuga, COMPSAC 2000)

Web Intelligence (WI)

Ning Zhong, Maebashi Inst. Tech.
Jiming Liu, Hong Kong Baptist U.
Y.Y. Yao, U. Regina
Setsuo Ohsuga, Waseda U.

1 Introduction

The 21st century is the age of Internet and World Wide Web. The Web revolutionizes the way we gather, process, and use information. At the same time, it also redefines the meanings and processes of business, commerce, marketing, finance, publishing, education, research, development, as well as other aspects of our daily life. The revolution is just beginning. Although individual Web-based information systems are constantly being deployed, advanced issues and techniques for developing and for benefiting from Web intelligence still remain to be systematically studied.

This position paper defines a new research field, namely *Web Intelligence* (WI for short) by giving a complete picture of WI related topics for systematic study on advanced Web technology and developing Web-based intelligent information systems. Roughly speaking, WI exploits AI and advanced information technology on the Web and Internet. It is the key and the most urgent research field of IT for business intelligence.

2 WI Related Topics

What are issues and research topics on WI? In order to study advanced Web technology systematically, and develop advanced Web-based intelligent information systems, we give an overview of WI related topics as shown in Figure 1 and list several major subtopics in each topic below.

- Web Human-Media Engineering:
  the art of Web page design, multimedia information representation, multimedia information processing, visualization of Web information, and Web-based human computer interface.
- Web Information Management:
  data quality management, information transformation, Internet and Web-based data management, multi-dimensional Web databases and (OLAP (on-line analytical processing)), multimedia information management, new data models for the Web, object oriented Web information management, personalized information management, semi-structured data management, use and management of metadata, Web knowledge management, Web page automatic generation and updating, as well as Web security, integrity, privacy and trust.
- Web Information Retrieval:
  approximate retrieval, conceptual information extraction, image retrieval, multi-linguistic information retrieval, multimedia retrieval, new retrieval models, ontology-based information retrieval, as well as automatic Web content cataloging and indexing.
- Web Agents:
  dynamics of information sources, e-mail filtering, e-mail semi-automatic reply, global information collecting, information filtering, navigation guides, recommender systems, remembrance agents, reputation mechanisms, resource intermediary and coordination mechanisms, as well as Web-based cooperative problem solving.
- Web Mining and Farming:
  data mining and knowledge discovery, hyperlink analysis and transformation, learning user profiles, multimedia data mining, regularities in Web surfing and Internet congestion, text mining, Web-based ontology engineering, Web-based reverse engineering, Web farming, Web-log mining, and Web warehousing.
- Web Information System Environment and Foundations:
  competitive dynamics of Web sites, emerging Web technology, network community formation and support, new Web information description and query languages, theories of small world Web, Web information system development tools, and Web protocols.
- Web-Based Applications:
  business intelligence, computational societies and markets, conversational systems, customer relationship management (CRM), direct marketing, electronic commerce and electronic business, electronic library, information markets, price dynamics and pricing algorithms, measuring and analyzing Web merchandising, Web-based decision support systems, Web-based

0-7695-0792-1/00 $10.00 © 2000 IEEE

Figure 1. A schematic diagram of WI related topics

distributed information systems, Web-based electronic data interchange (EDI), Web marketing and Web publishing

3 WI Related Case Studies

WI presents an excellent opportunity as well as challenge to the research and development of new generation of information processing technology, as well as for exploiting business intelligence. Specifically, e-commerce activity that involves the end user is undergoing a significant revolution [8]. The ability to track users' browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before. It is now possible for a vendor to personalize his product message for individual customers at a massive scale. This is called *Targeted Marketing*.

Hackathorn proposed Web farming that is the systematic refining of information resources on the Web for business intelligence [3].

Ahlstrand et al. systematically investigated the data on the Web and the features of semi-structured data [1].

Zhong, Yao et al. proposed a way of mining peculiar data and peculiarity rules that can be used for Web-log mining [11]. They proposed ways for targeted marketing by using classification rules and concept value functions [15, 10]. They are also working on text mining on the Web including automatic construction of ontology, e-mail filtering systems, and Web-based e-business systems [14, 16].

Liu et al. are working on e-commerce agents [7]. Liu and Zhong are also working on Web agents and EDDIA (Knowledge Discovery and Data Mining Agents) [5, 6].

4 WI Conferences

We initiated a new high-quality, high-impact biennial conference series, namely the Asia-Pacific Conference on Web Intelligence (WI). The first meeting in this new series, WI'01, will be held in Maebashi City, Japan, October 23-26, 2001 (http://kis.maebashi-it.ac.jp/wi2001).

WI'2001 is an international forum for researchers and practitioners to present the state-of-the-art in the development of Web intelligence, to examine performance characteristics of various approaches in Web-based intelligent information technology, and to cross-fertilize ideas on the development of Web-based intelligent information systems among different domains. By idea-sharing and discussions on the underlying foundations and the enabling techniques of Web intelligence, WI'2001 is expected to stimulate the future development of new models, new methodologies, and new tools for building a variety of embodiments of Web-based intelligent information systems.

References

[1] Ahlstrand, S., Isaacson, P. and Stein, J. *Data on the Web*. Morgan Kaufmann (2000).

[2] Hong, J.Z., Zhong, N., an Ohsuga, S. "Probabilistic Rough Induction: The GDT-RS Methodology an Algorithms", Z.W. Ras and A. Skowron (eds.), *Foundations of Intelligent Systems*, LNAI 1609, Springer-Verlag (1999) 621-628

[3] Hackathorn, RD. *Web Farming for the Data Warehouse*. Morgan Kaufmann (2001).

[4] Liu, C.L. and Li, C. "Data Mining for Stock Marketing: problems and solutions", *Proceedings of IEEE*, 88 (1998) 70-78

[5] Liu, J. and Zhong, N. (eds.) *INTELLIGENT AGENT TECHNOLOGY: Systems, Methodologies and Tools*. World Scientific (1999).

[6] Liu, J., Zhong, N., Yao, Y.Y. and Ras, Z.W. "The Wisdom Web: New Challenges for Intelligence", *Journal of Intelligent Information Systems*, World Scientific (2001).

[7] Liu, J. and Ye Y. (eds.) *Advances in E-commerce Agents: Brokerage, Negotiation, Security, and Mobility*. Springer-Verlag (2001).

[8] Srivastava, J. et al. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations Newsletter of SIGKDD, Vol.1 Issue 2 (2000) 12-23.

[9] Yao, Y.Y. and Dong, N. "Maximal Mobile Value Functions for Targeted Marketing" (manuscript).

[10] Yao, Y.Y. and Zhong, N. "Granular Computing using Information Tables". Lin, T.Y., Yao, Y.Y. and Zadeh, L.A. (eds.) *Data Mining, Rough Sets and Granular Computing*, Physica-Verlag (2001).

[11] Zhong, N., Yao, Y.Y., and Ohsuga, S. "Peculiarity Oriented Multi-Database Mining", J. Zytkow and J. Rauch (eds.) *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial Intelligence 1704, Springer-Verlag (1999) 136-146.

[12] Zhong, N., Dong, J.Z., and Ohsuga, S. "Data Mining: A Probabilistic Rough Set Approach", in Polkowski, L., Tsumoto, S. and Lin, T.Y. (eds.) *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, Physica-Verlag (2000).

[13] Zhong, N. and Zhou, L. (eds.) *Methodologies for Knowledge Discovery and Data Mining*, LNAI 1574, Springer-Verlag (1999).

[14] Zhong, N., Yao, Y.Y., and Kakemoto, Y. "Automatic Construction of Ontology from Text Databases", *Proc. Second International Conference on DASFAA'99*, World Scientific Press (1999).

[15] Zhong, N., Dong, J.Z., and Ohsuga, S. "Rule Discovery by Soft Induction Techniques", in Special Issue on Rough-Neuro Computing, *Neurocomputing: An International Journal*, Elsevier Science Publishers (2000).

[16] Zhong, N. "A Study on E-mail Filtering in Uncertainty Environment: Automatic Rule Learning", Technical Report, Yamaguchi University (2000).

[17] Zhong, N. *Knowledge Discovery and Data Mining*, in the Encyclopedia of Microcomputers, Volume 27 (Supplement 6), Marcel Dekker, Inc. (2001).

Une brève histoire du Web Intelligence

- 2001: First Asia-Pacific Conference on Web Intelligence
- 2002: Publication of first special issue on WI in IEEE Computer
- 2002: Web Intelligence Consortium
- 2003: First edited book on WI
- 2005: The international WIC Institute

Motivations

- **La taille de Web**
  Difficultés dans le stockage, la gestion, et la récupération (recherche) effective et efficace

- **Complexité du Web**
  - Collection hétérogène de documents web structurés, non structurés, semi-structurés, interdépendants et distribués.
  - Consiste en textes, images et sons

Motivations

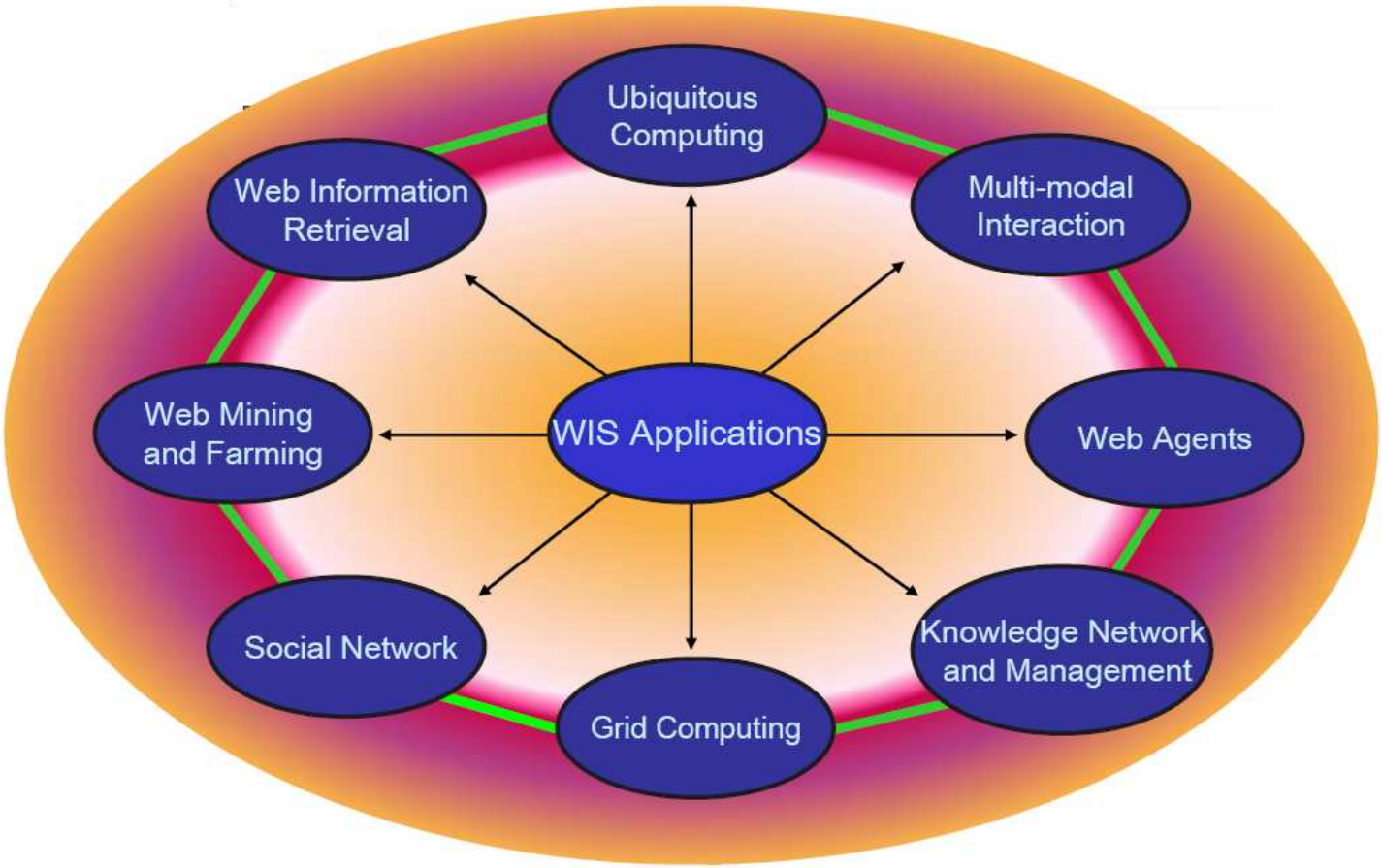
Web Intelligence on the Web

Search Engine	Number of Hits (February 2001)	Number of Hits (February 2003)
Lycos	1,102,279	**7,163,922**
Google	1,080,000	**2,590,000**
AltaVista	1,271	**1,860,062**
Netscape	77	**1,900,000**
Yahoo	74	**2,510,000**

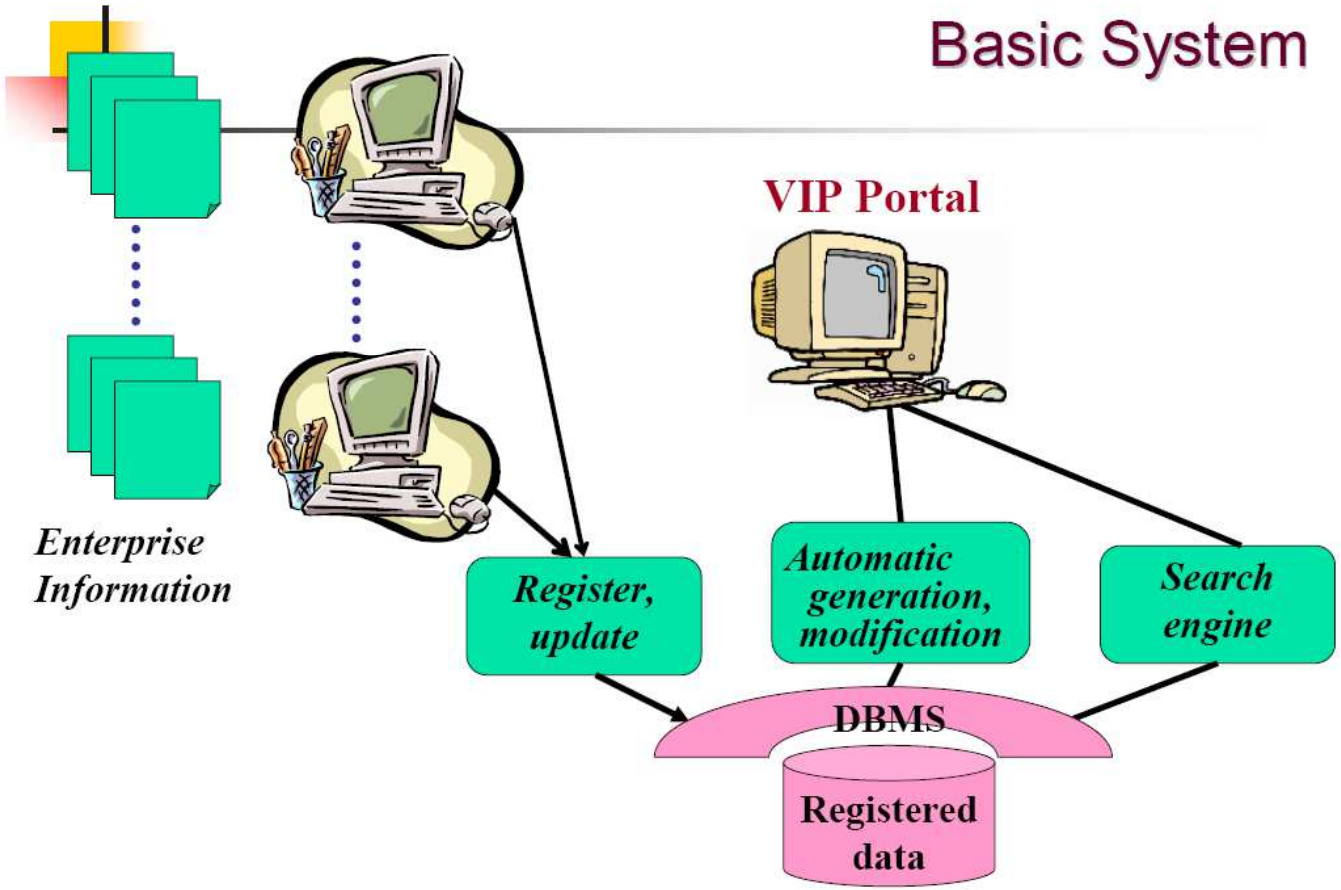
Motivations

- Production de données sur le Web est à un taux de croissance exponentielle.
- Un intérêt industriel en croissance rapide dans WI.
- Seuls quelques articles académiques.
- Nous devons réduire l'écart entre les besoins de l'industrie et la recherche universitaire

Système en Web Intelligence



Un exemple


Basic System
VIP Portal
Enterprise Information
Register, update
Automatic generation, modification
Search engine
DBMS
Registered data

Questions avancée

- Comment le client utilise le portail VIP afin de cibler les produits et gérer les promotions et campagnes marketing?
- Quel est l'association sémantique entre les pages visitées par le client?
- Est-ce que le visiteur est familiariser avec la structure du site Web? Ou est-il un nouvel utilisateur ou quelqu'un au hasard?
- Est-ce que le visiteur est un robot Web ou d'autres utilisateurs?
- …

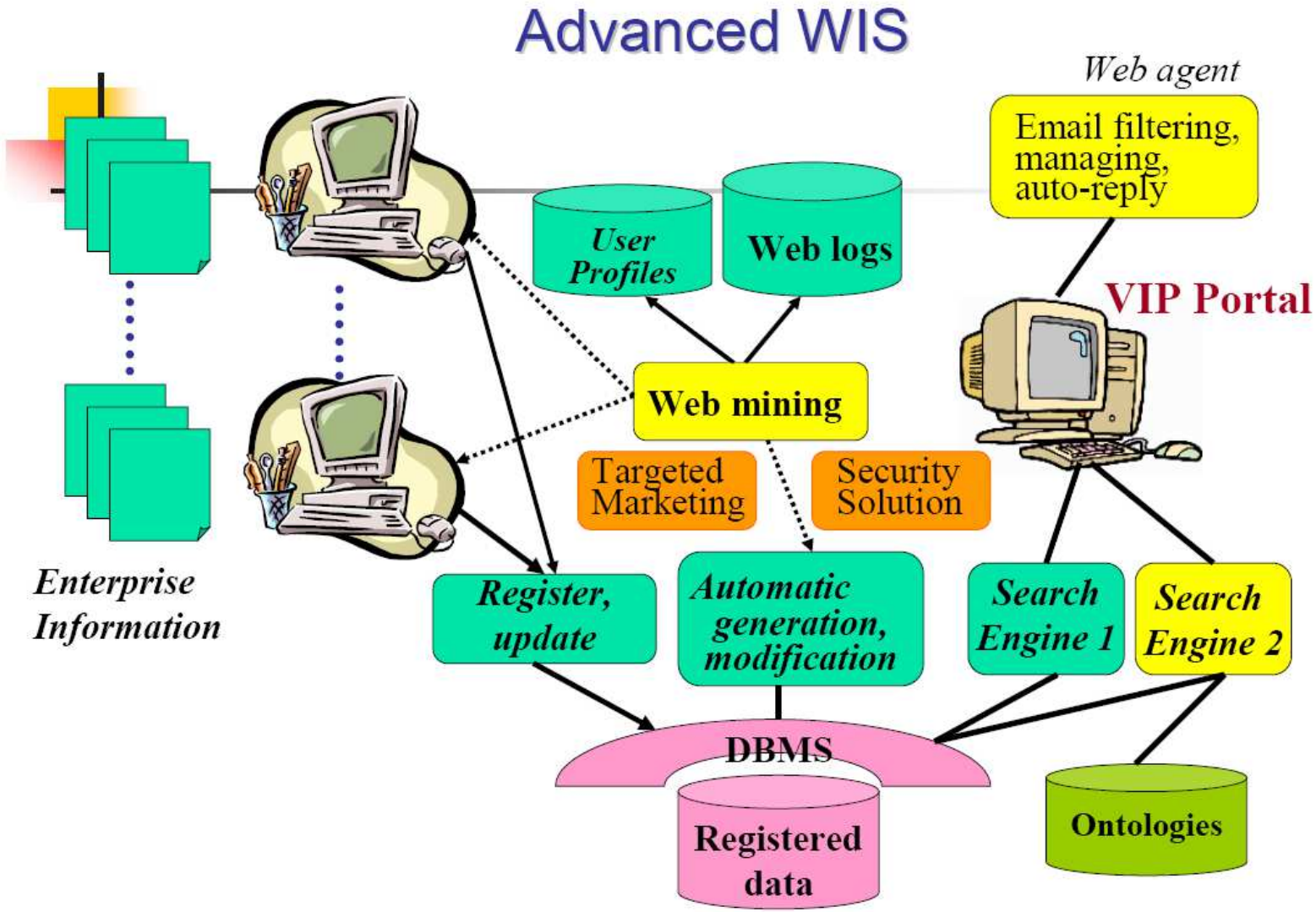
système WI avancé

- Faire une recommandation dynamique à un utilisateur Web basé sur le profil de l'utilisateur et les comportements d'utilisation;

- La modification automatique du contenu et de l'organisation d'un site web;

- En combinant les données d'utilisation sur le Web avec des données marketing pour donner des informations sur la façon dont les visiteurs ont utilisé un site Web.

Système avancé pour le WI



Perspectives du WI

- WI peut être classés en quatre catégories
- (based on Russel & Norvig`s scheme)

conception de la philosophie de WIS

capacité,
la fonctionnalité
d'un WIS

System that **thinks** like **humans**	System that **thinks** **rationally**
System that **acts** like **humans**	System that **acts** **rationally**

Intérêts industriels dans WI

- **Web Intelligence** kis-lab.com/wi01/
- **Web-Intelligence** Home Page
  - **www.web-intelligence.com/**
- **Intelligence** on the **Web**
  - **w**ww.fas.org/irp/intelwww.html
- WIN: home **WEB INTELLIGENCE** NETWORK,
  - smarter.net/
- CatchTheWeb - **Web** Research, **Web Intelligence** Collaboration www.catchtheweb.com/
- Infonoia: **Web Intelligence** In Your Hands
  - www.infonoia.com/myagent/en/baseframe.html

Programme de recherche sur le WI

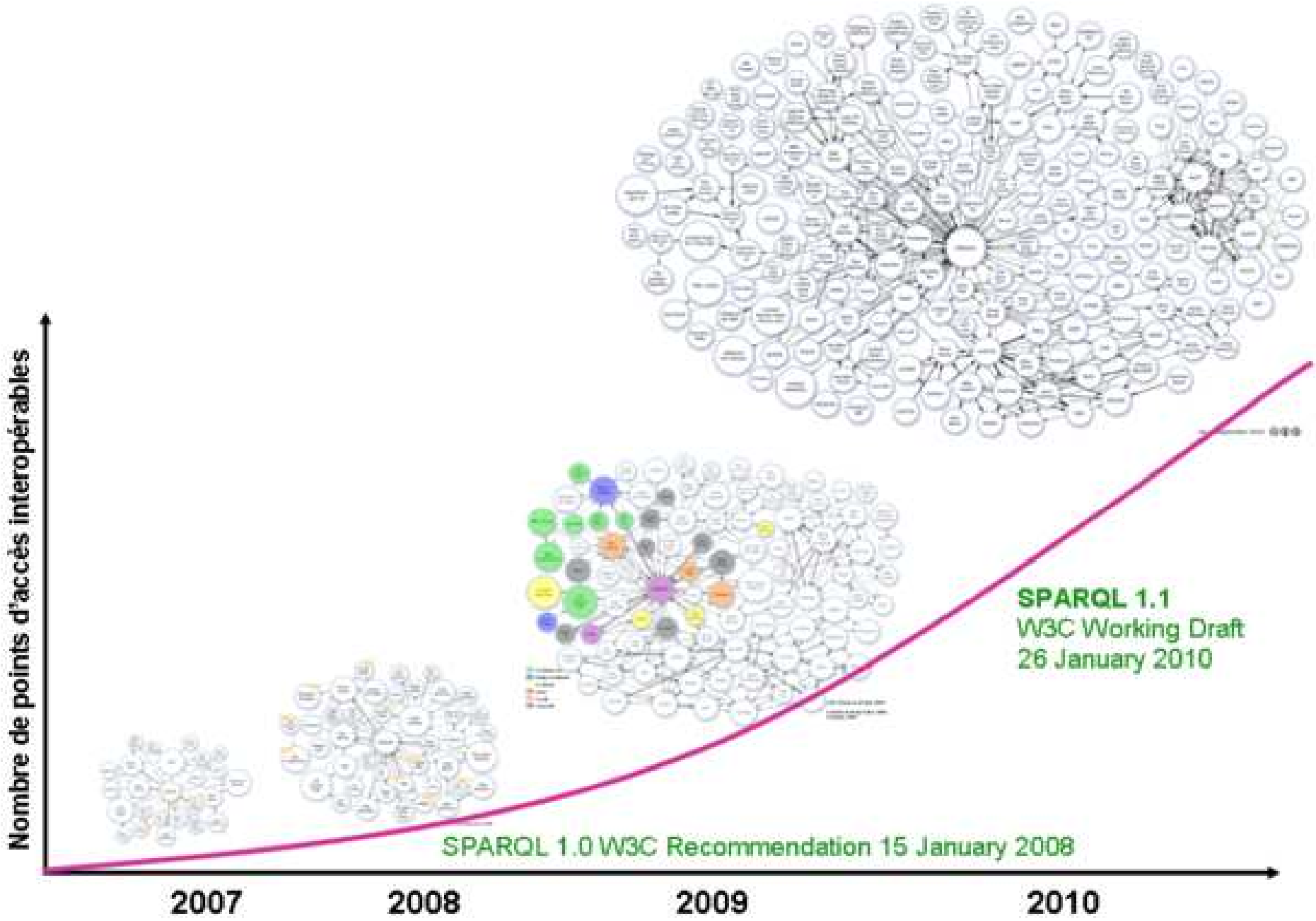
- Web sémantique, « webmining » et construction automatique d'ontologies
- réseau social intelligent

- Cloud computing…
- Application : e-business, e-learning, e-santé, (mobile, ubiquitous….)
- …

Le Web Sémantique (3.0)

- Le Web sémantique est basé sur les languages qui font plus de contenu sémantique de la page disponible en formats lisibles par la machine pour un traitement automatique basé sur des agents.

- Un langage «sémantique» qui lie les informations sur une page à la sémantique lisibles par la machine s'appelle « ontologie ».



Le **Web** Sémantique ou l'importance des données liées

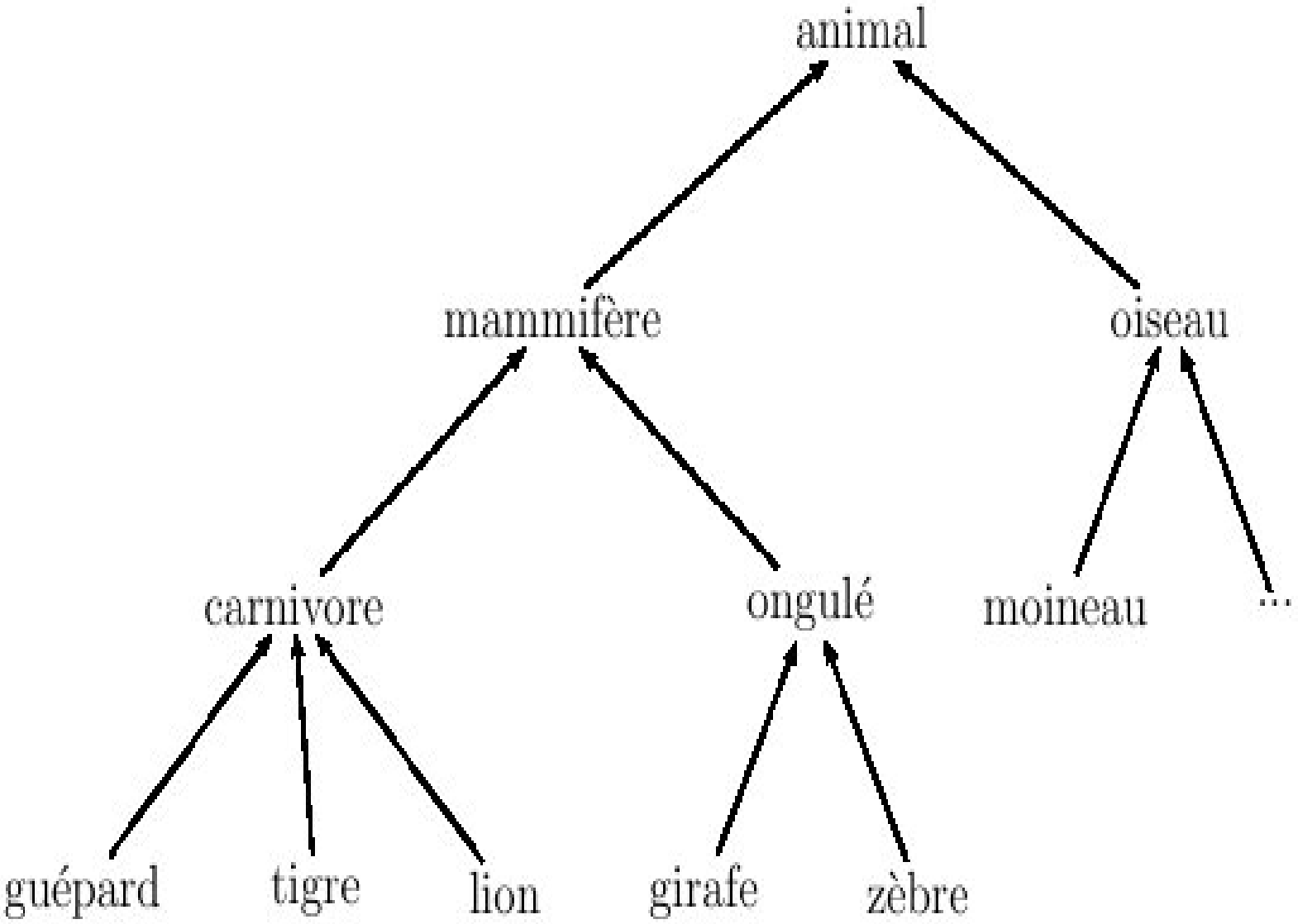
Compoants du Web Semantique

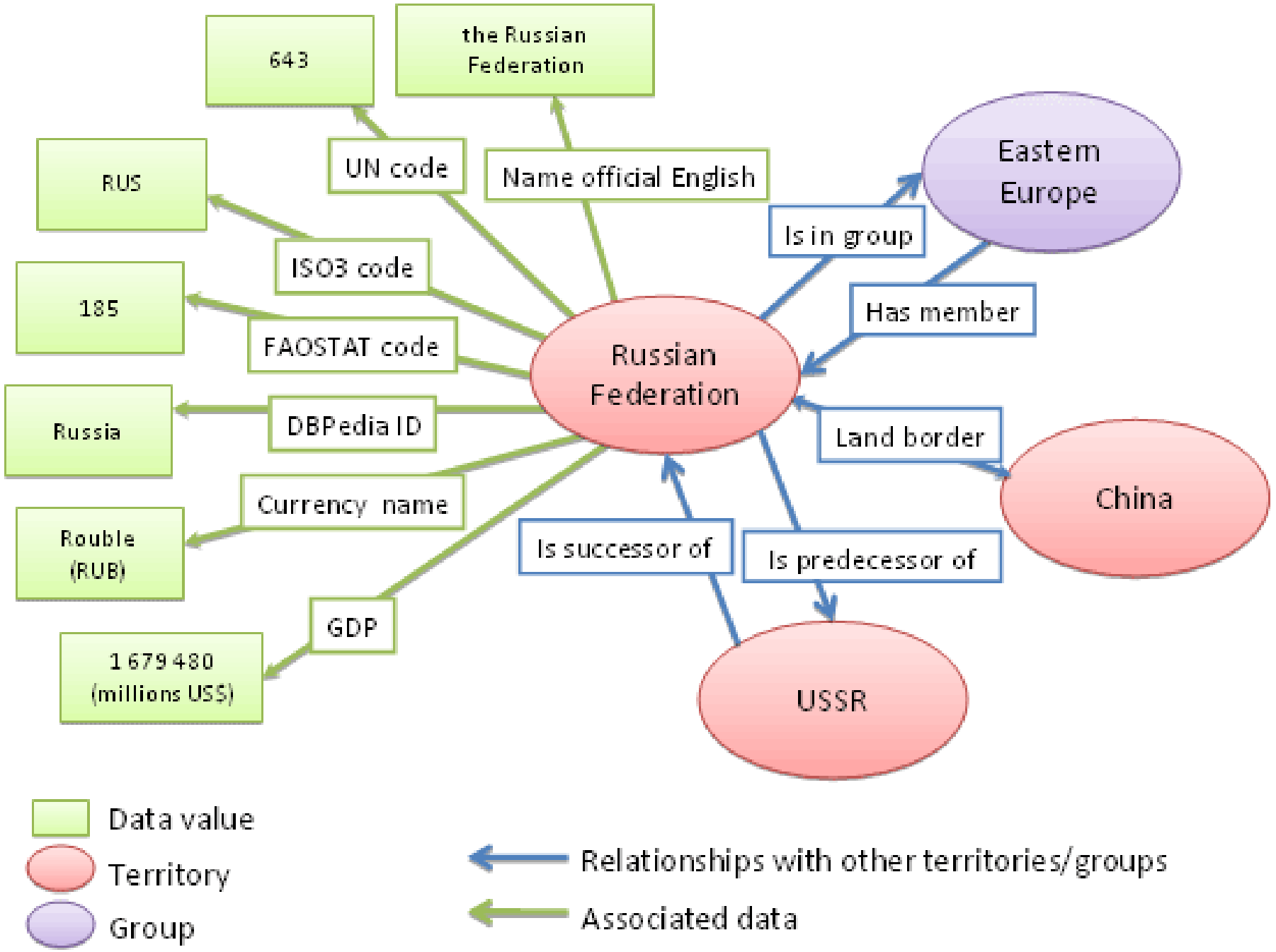
- Un modèle de données unifié comme RDF.
- Langues avec sémantique définie, construit sur RDF, tels que OWL (DAML + OIL).
- Ontologies pour normaliser la terminologie pour le balisage des ressources Web.
- Les outils qui aident la génération et le traitement de balisage sémantique.

Les Ontologies fournissent l'épine dorsale sémantique pour les applications du Web sémantique.

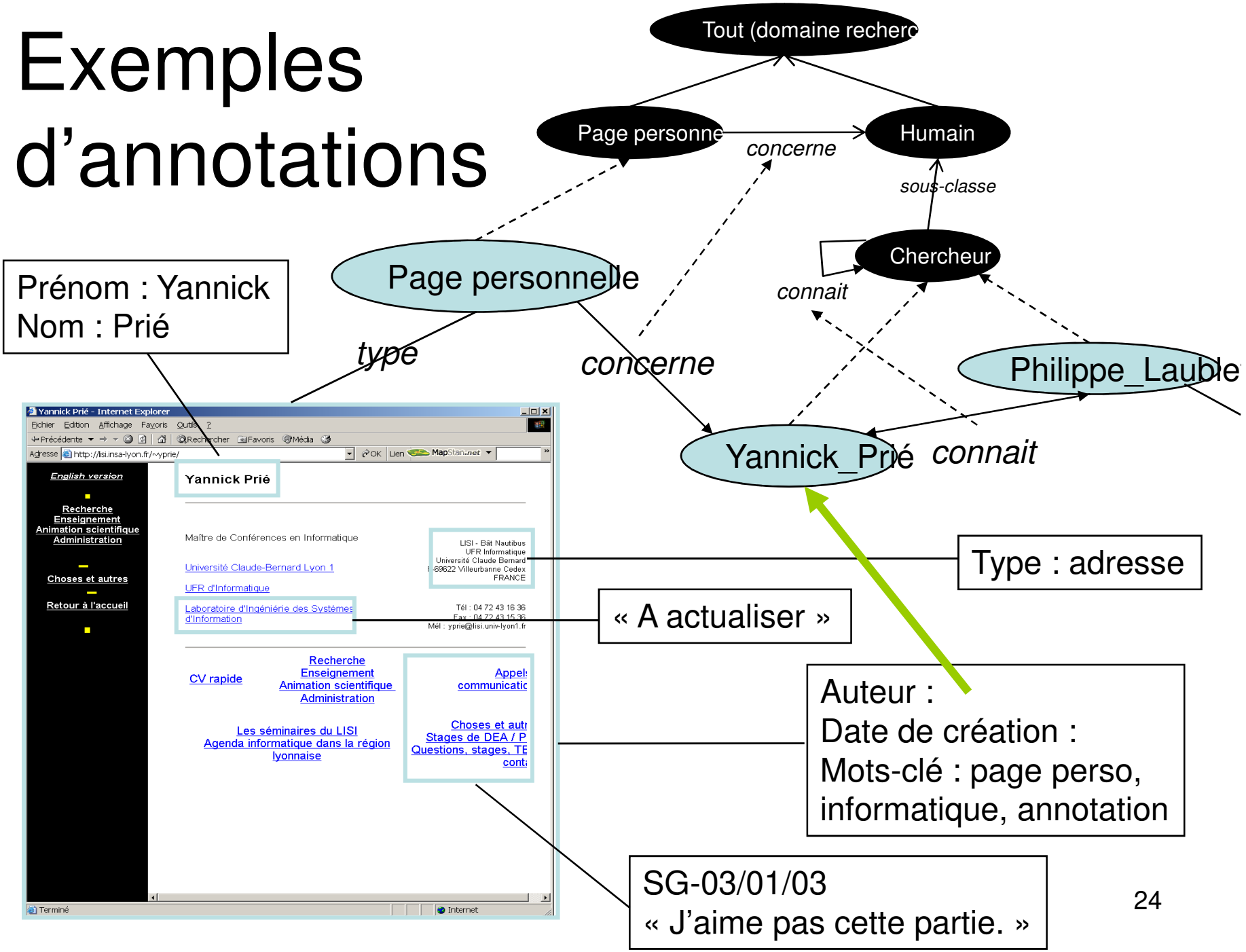
Les ontologies offrent

- Communication
  - Les modèles normatifs, des réseaux de relations
- Partage et réutilisation
  - Spécifications, Fiabilité
- Contrôle
  - Classification et découverte, partage, découverte des relations


animal
mammifère
oiseau
carnivore
ongulé
moineau
...
guépard
tigre
lion
girafe
zèbre


643
the Russian Federation
UN code
Name official English
RUS
ISO3 code
185
FAOSTAT code
Russia
DBPedia ID
Currency name
Rouble (RUB)
GDP
1 679 480 (millions US$)
Russian Federation
Is in group
Eastern Europe
Has member
Land border
China
Is successor of
Is predecessor of
USSR
Data value
Territory
Group
Relationships with other territories/groups
Associated data

Exemples d'annotations



Categories des Ontologies

- Une ontologie spécifique au domaine décrit un domaine technique ou commercial bien défini.

- Une ontologie de tâche peut être soit spécifique à un domaine ou reconstituées à partir d'un ensemble d'ontologies spécifiques au domaine pour répondre aux exigences d'une tâche.

- Une ontologie universelle décrit connaissances à des niveaux supérieurs

Programme de recherche sur le WI

- Web sémantique, « webmining » et construction automatique d'ontologies
- réseau social intelligent

- Cloud computing…
- …
- …

Social Network

- Adopter et intégrer le web 2.0 sont plus faciles que le définir
- Web 1.0 vs Web 2.0

Facteurs de comparaison	Web 1.0	Web 2.0
Contribution aux contenus	Dépend du Webmestre	Dépend des internautes
Fréquence actualisation	Relativement élevée	Très élevée (participation active d'un grand nombre d'intervenants)
Relation	« **Un à plusieurs** » : webmestre aux lecteurs	« **Plusieurs à plusieurs** » : l'internaute est à la fois lecteur et auteur (peut réagir, interagir...)

- Principe du « push » et du « pull »
- Pas un effet de mode, mais un véritable phénomène :
  - Personnalité de l'année 2006 magazine Time : « YOU » (Vous, les internautes générant le contenu sur Internet)
  - Loin d'être éphémère

Le Web comme un Graphe

- Nous pouvons voir le Web comme un réseau social orienté qui relie les gens (organisations ou entités sociales).

- Questions de recherche:

  - Quelle est la taille du graphe? (degré sortant et degré entrant)

  - Peut-on naviguer à partir de n'importe quelle page à une autre? (clics)

  - Pouvons-nous exploiter la structure du Web? (recherche et webmining)

  - Comment découvrir et gérer les communautés Web?

  - Est-ce que le graphe du Web révèle sur la dynamique sociale?

Outline

- A brief history of Web Intelligence
- Motivations for WI
- Definition and Perspectives of WI
- Agenda de Recherche
- **Les outils Majeurs pour le Web Intelligence**
- Conclusion

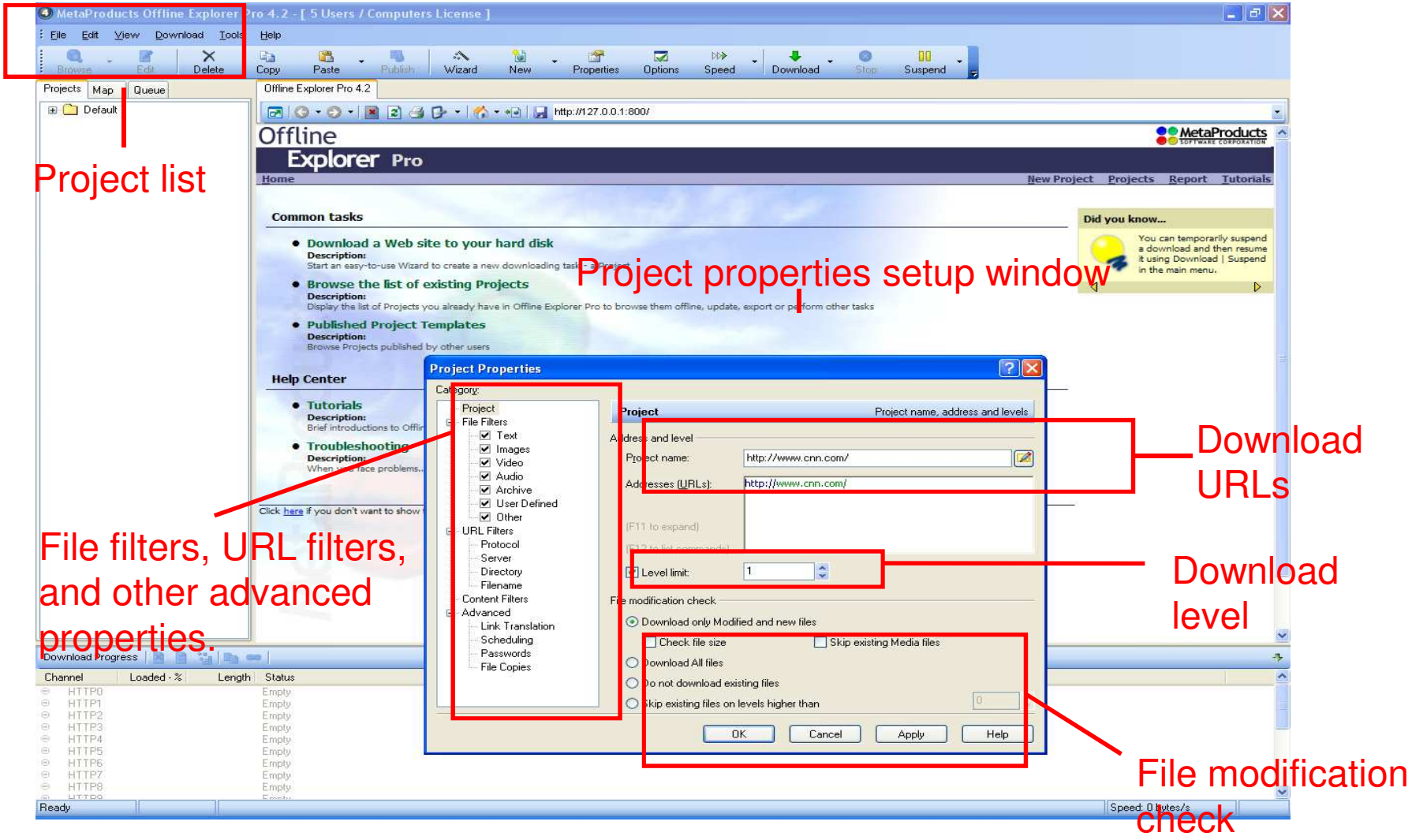
Principaux outils de Web Intelligence

- I. Collection
  - Offline Explorer
  - SpidersRUs (AI Lab)
  - Google Scholar

- II. Analyses (Données et Text Mining)
  - Google APIs
  - Google Translation
  - GATE
  - Arizona Noun Phraser (AI Lab)
  - Self-Organizing Map, SOM (AI Lab)

- III. Visualisation
  - NetDraw
  - JUNG
  - Analyst's Notebook and Starlight

Collection: Offline Explorer



• Offline Explorer télécharge vos sites Web et FTP favoris (jusqu'à une centaine simultanément) afin de les consulter hors connexion, de les imprimer ou de les parcourir.

• Une des forces de Offline Explorer est de vous donner accès de manière sélective aux différents serveurs (inclus ou exclus), répertoires et fichiers en utilisant seulement des mots-clé.

• Efficace et rapide, il comprend un support de technologies au standard industriel, comme FTP, différents serveurs proxy, Macromedia Flash, Cookies, un serveur HTTP intégré vous permettant de partager vos fichiers téléchargés sur l'Intranet.

Analysis: Google APIs

- L'*API Google* (application programming interface) est un ensemble d'outils mis à disposition de tous qui permet d'interroger à distance les serveurs du Moteur de recherches *Google*.
  http://code.google.com/more/

- Examples of Google APIs:
  - Google API for Inlink: Discovers what pages link to your website.
  - Google Data APIs: Provide a simple, standard protocol for reading and writing data on the Web. Several Google services provide a Google Data API, including Google Base, Blogger, Google Calendar, Google Spreadsheets and Picasa Web Albums.
  - Google AJAX Search API: Uses JavaScript to embed a simple, dynamic Google search box and display search results in your own Web pages.
  - Google Analytics: Allows users gather, view, and analyze data about their Website traffic. Users can see which content gets the most visits, average page views and time on site for visits.
  - Google Safe Browsing APIs: Allow client applications to check URLs against Google's constantly-updated blacklists of suspected phishing and malware pages.
  - YouTube Data API: Integrates online videos from YouTube into your applications.

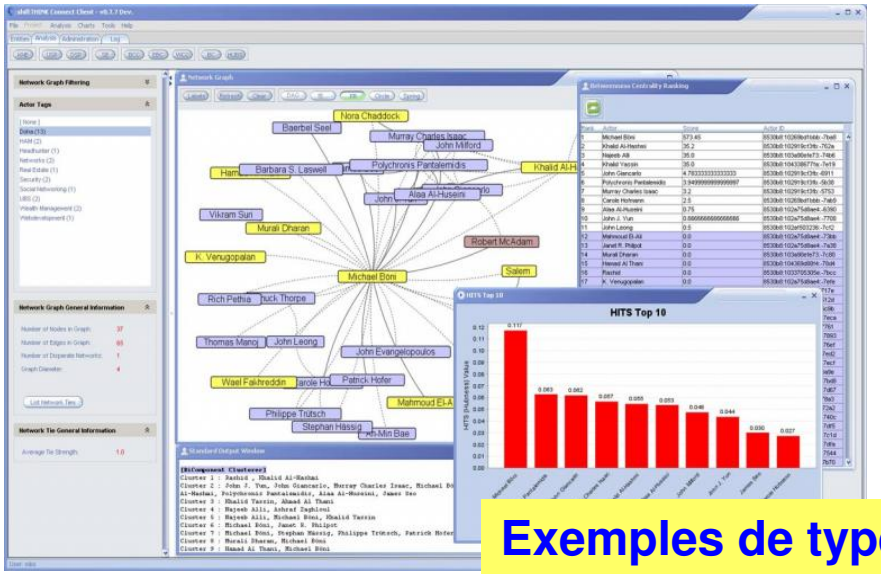
Visualization: JUNG

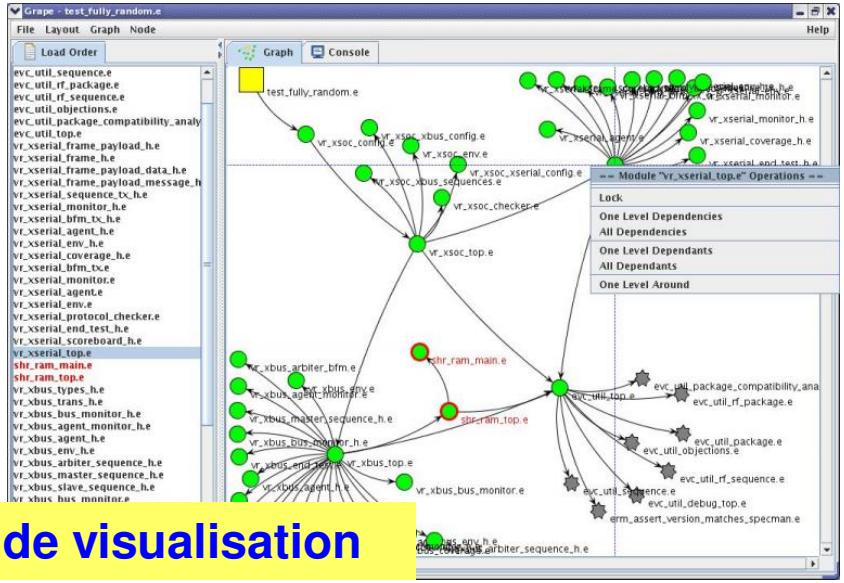
- Java Universal Network/Graph Framework (JUNG) est une bibliothèque de logiciels pour la modélisation, l'analyse et la visualisation des données qui peuvent être représentéees sous forme de graphique ou de réseau. Il a été développé par l'École de l'information et d'informatique à l'Université de Californie, Irvine

  http://jung.sourceforge.net/index.html

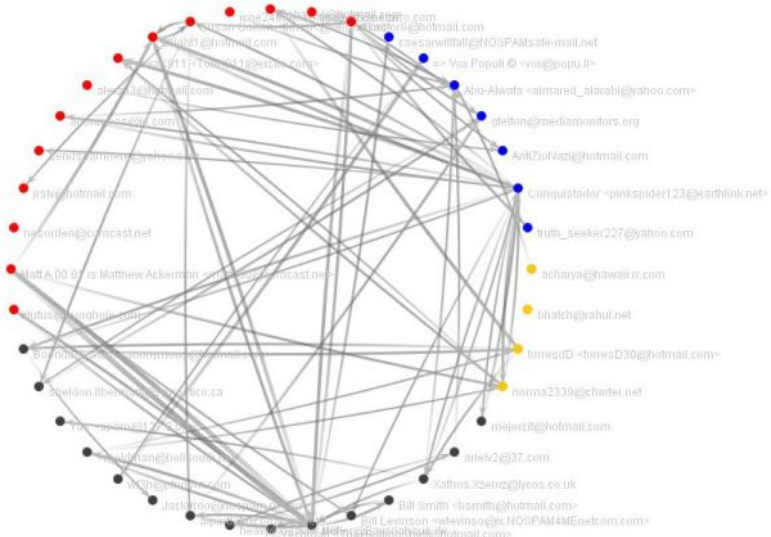
- La distribution actuelle de JUNG comprend des implémentations d'un certain nombre d'algorithmes de la théorie des graphes, data mining et d'analyse de réseau social
  - Clustering
  - Decomposition
  - Optimization
  - Random Graph Generation
  - Statistical Analysis
  - Calculation of Network Distances and Flows and Importance Measures (Centrality, PageRank, HITS, etc.).

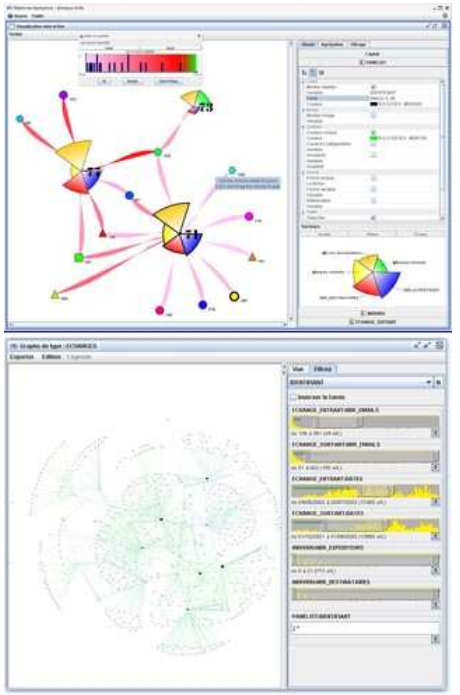
JUNG





Exemples de types de visualisation





A retenir

- Le couplage de l'hypertexte d'Internet conduit à une révolution: le Web.

- Le couplage de l'intelligence artificielle et de la technologie d'information avancé sur la plate-forme de Web, mènera à un autre changement de paradigme: le Web Intelligent.

.

Fin du premier cours