Dynamic KOS building & management for library information systems

Piotr Gawrysiak, Henryk Rybiński
{gawrysia, rybinski}@ii.pw.edu.pl
Warsaw University of Technology

Michał Okoniewski
michal.okoniewski@fao.org
Food & Agriculture Organization of the UN

Presented by **Stefka Kaloyanova**, FAO of the UN

16th September 2004

Presentation overview

1. SEMKOS project overview & follow-up
2. Text Mining and Ontologies
3. Text Mining system overview
4. FAO library systems integration
5. Conclusions & future work

SEMKOS project

SEMKOS = *Semantic enabling by advanced Knowledge Organization Systems for large scale information integration in scientific and cultural digital libraries*

SEMKOS Project proposal has been submitted to EC Commission under Sixth Framework Programme

SEMKOS consortium that was established included Warsaw University of Technology as a R&D centre and FAO as a testbed institution

The commission rejected our proposal, but WUT team decided to not disband itself and continue working in loose collaboration with FAO

Final goal that we want to achieve

↓

Construction of a specialized, ontology building & library management oriented text mining system

Supporting ontology systems with TM

There are many opinions about what is the main obstacle to building usable ontology systems. Issues often mentioned include:

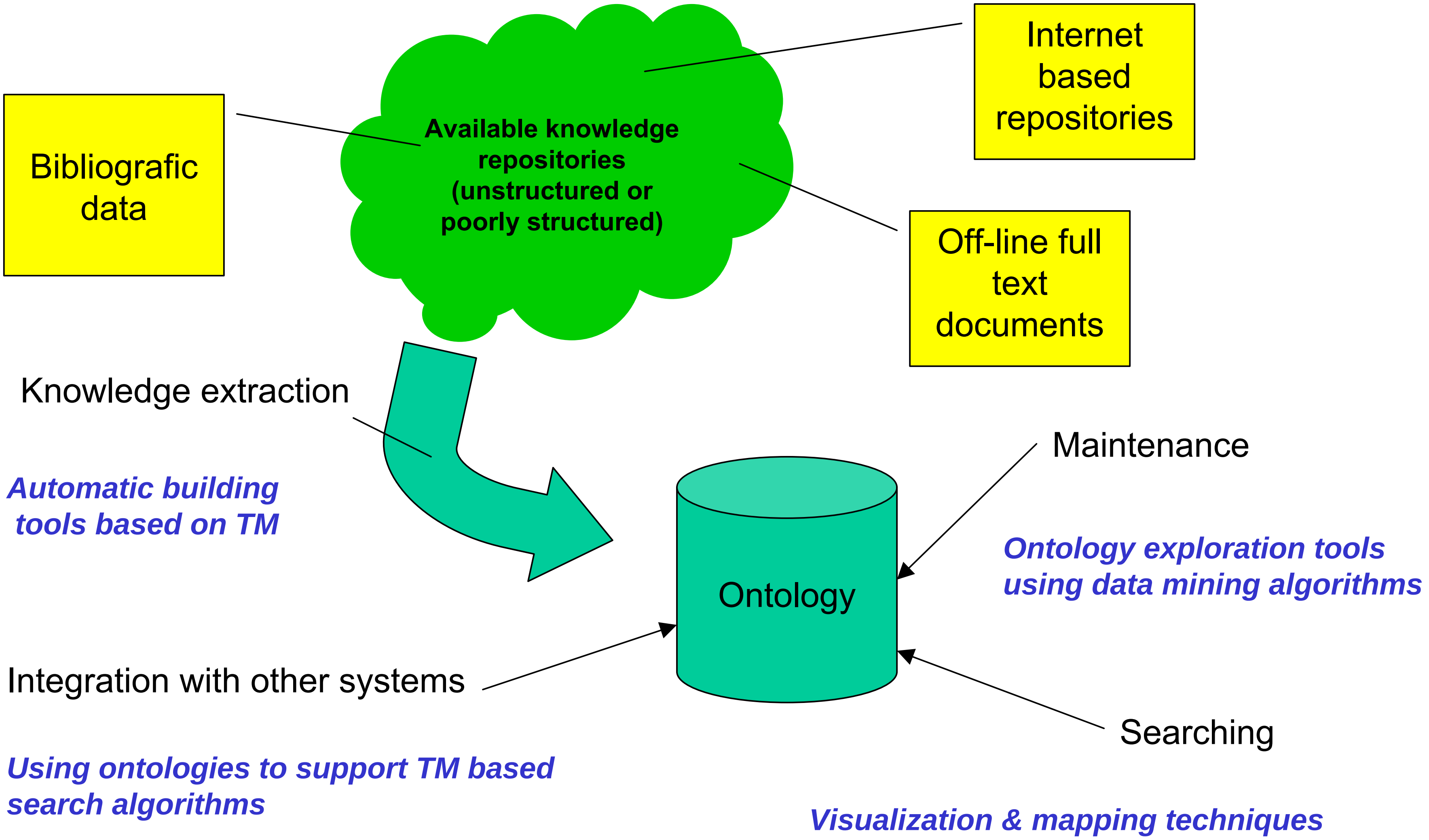
- lack of data & difficulties in transforming loosely structured, existing knowledge systems, into rigid ontologies
- often poor quality of newly constructed ontologies
- problems with applying ontologies in real-world systems
    - *integration problems*
    - *data searching problems*
    - *ontology merging problems*
    - *...*

Ontologies design & creation is a objectively <u>difficult</u> task, not well handled by <u>human editors</u>

Note, that there are no equivalents of ontology knowledge representation systems in nature – all bio-systems store & process knowledge represented in unstructured form

As TM is a method of extracting useful and rigid knowledge from huge amounts of unstructured repositories, it can probably help here

TM applications in KOS



Available knowledge repositories (unstructured or poorly structured)

Bibliografic data

Internet based repositories

Off-line full text documents

Knowledge extraction

Automatic building tools based on TM

Ontology

Maintenance

Ontology exploration tools using data mining algorithms

Integration with other systems

Using ontologies to support TM based search algorithms

Searching

Visualization & mapping techniques

„Canonical” TM methods

- document clustering
- document classification
- keyword and keyphrase identification
- document summarization
- automatic language identification
- parts-of-speech tagging
- document repository visualization
- automatic language translation

Metadata quality problem

Many standard Text Mining approaches assume that the corpus contains valid (in a grammatical and orthographical context) contents.

Alas, in many library systems the bibliographical texts are often very short and contain high amount of noise:

- spelling mistakes
- typos
- incorrect
- incosistent classification hierarchies
- ...

Also, there is often no verification of metadata in many systems and in turn its quality tends to be poor due to manual manipulation.

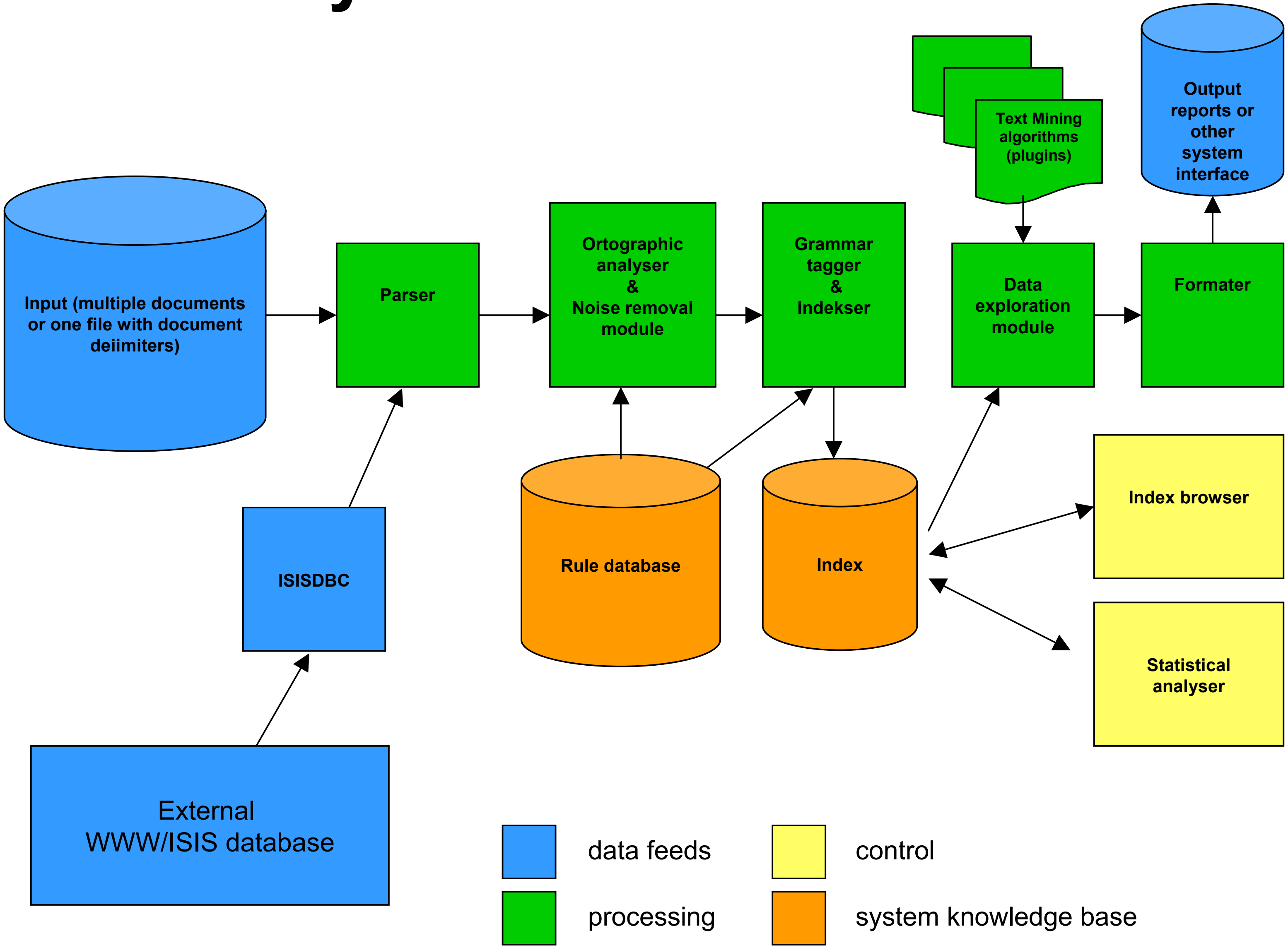
Above problems make searching and analysing metadata very difficult task

We need a system that would be at least partially immune to these problems

↓

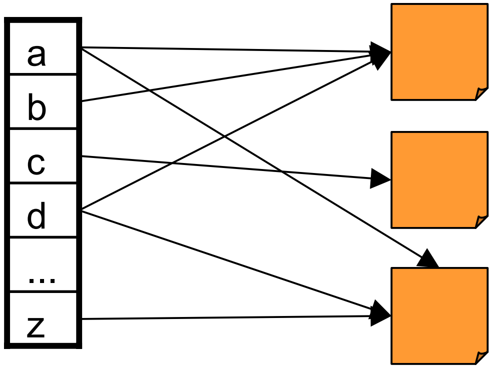
XXX Text Mining System

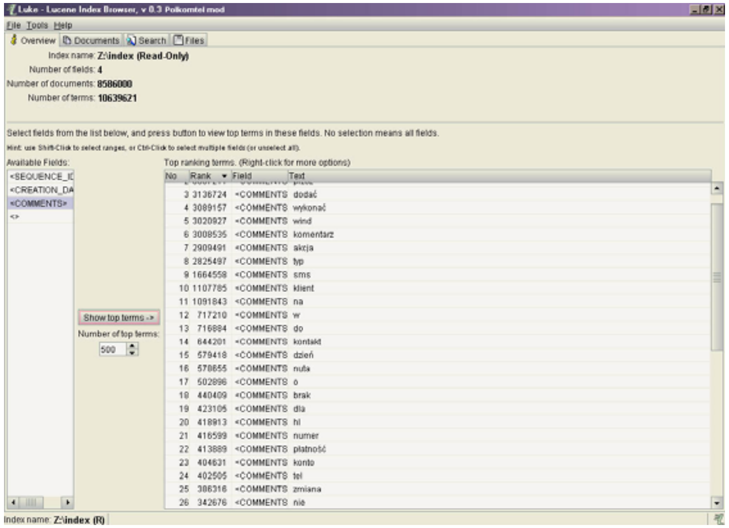
XXX TM system structure


Text Mining algorithms (plugins)
Output reports or other system interface
Input (multiple documents or one file with document deiimiters)
Parser
Ortographic analyser & Noise removal module
Grammar tagger & Indekser
Data exploration module
Formater
ISISDBC
Rule database
Index
Index browser
Statistical analyser
External WWW/ISIS database
data feeds
control
processing
system knowledge base

System properties

- **Implementation language** - Java J2EE 1.4
- **Main index** – an extension of *Apache Lucene* project, both forward and reverse index have been implemented
- **Modular and extensible** – individual TM algorithms can be easily „plugged in”
- **Web application integration** – *Jakarta* compatible, can be linked with *Labeo/Turbine* platform applications
- **Text noise management capabilities** – able to process text that contains high amount of orthographic (due to typos or **OCR process**) and grammar errors
- **Efficiency** –*able to filter, index & correct 1mln 1 sentence texts* / *1h* on a standard office PC, several orders of magnitute faster on AIX mainframe





Index name: Z:\index (Read-Only)
Number of fields: 4
Number of documents: 8586000
Number of terms: 10639621

Select fields from the list below, and press button to view top terms in these fields. No selection means all fields.

Hint: use Shift-Click to select ranges, or Ctrl-Click to select multiple fields (or unselect all).

Available Fields:
<SEQUENCE_ID
<CREATION_DA
<COMMENTS>
<>

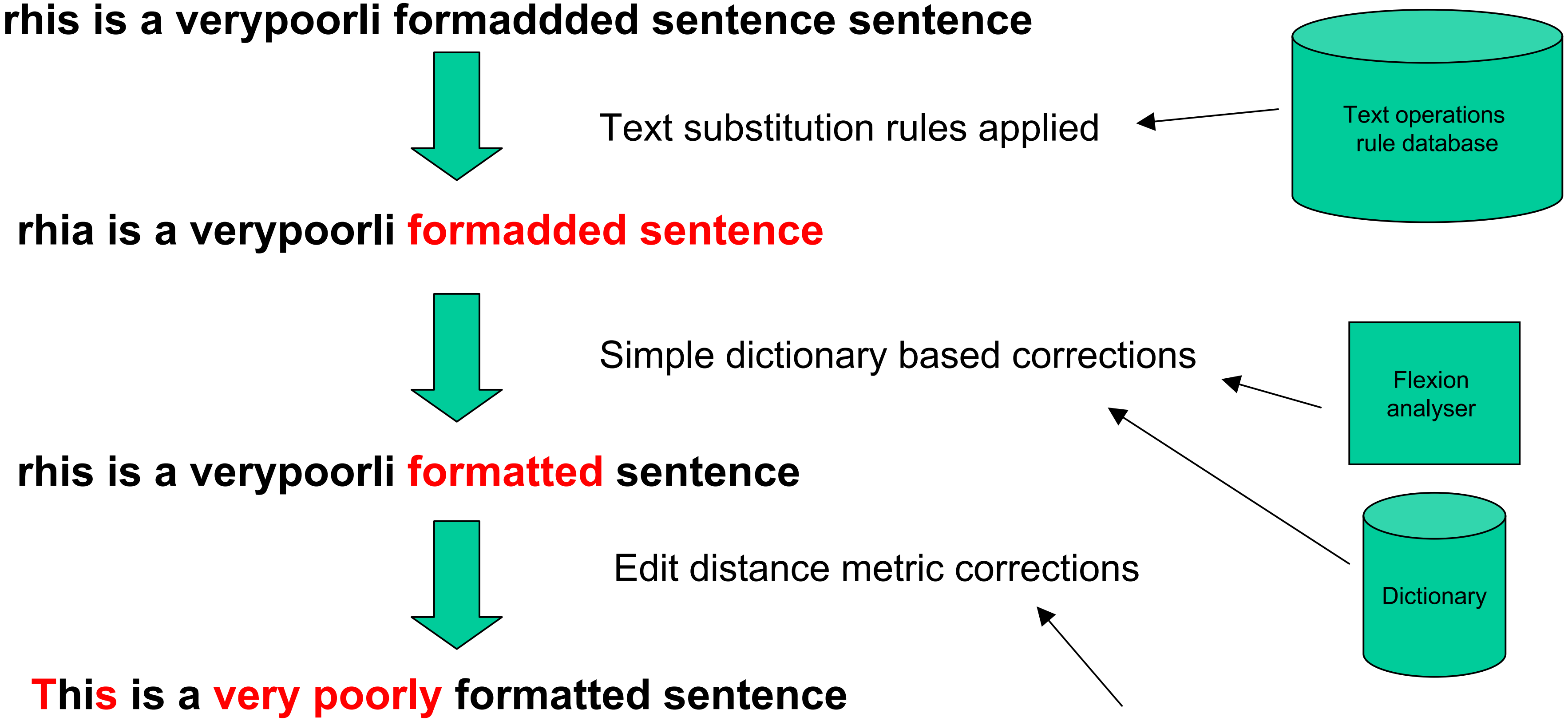
Number of top terms: 500

Top ranking terms. (Right-click for more options)

No	Rank	Field	Text
3	3136724	<COMMENTS	dodać
4	3089157	<COMMENTS	wykonać
5	3020927	<COMMENTS	wind
6	3008535	<COMMENTS	komentarz
7	2909491	<COMMENTS	akcja
8	2825497	<COMMENTS	typ
9	1664558	<COMMENTS	sms
10	1107785	<COMMENTS	klient
11	1091843	<COMMENTS	na
12	717210	<COMMENTS	w
13	716884	<COMMENTS	do
14	644201	<COMMENTS	kontakt
15	579418	<COMMENTS	dzień
16	570655	<COMMENTS	nula
17	502896	<COMMENTS	o
18	440409	<COMMENTS	brak
19	423105	<COMMENTS	dla
20	418913	<COMMENTS	hl
21	416599	<COMMENTS	numer
22	413889	<COMMENTS	płatność
23	404631	<COMMENTS	konto
24	402505	<COMMENTS	tel
25	386316	<COMMENTS	zmiana
26	342676	<COMMENTS	nie

Index name: Z:\index (R)

Noise management



rhis is a verypoorli formaddded sentence sentence

Text substitution rules applied

Text operations rule database

rhia is a verypoorli formadded sentence

Simple dictionary based corrections

Flexion analyser

rhis is a verypoorli formatted sentence

Edit distance metric corrections

Dictionary

This is a very poorly formatted sentence

Cost function based on:

- Keyboard layout
- Common spelling mistakes database
- Dictionary analysis

Other parsing operations:

- ***synonym replacement using thesaurus database***
  - *English language – WordNet based*
  - *Polish language – proprietary*
- ***abbreviation analysis***
- ***key phrases analysis** (example - temporal key phrases: „on Saturdays” -> time_day_of_week*

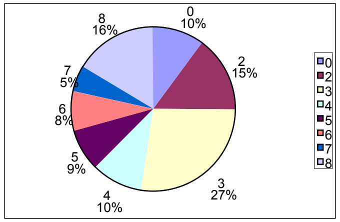
Business experiments – system verification

XXX system has been used in a customer relationship management analysis for Polkomtel S.A. (a GSM operator) in a project undertaken by Institute of Computer Science, Warsaw University of Technology

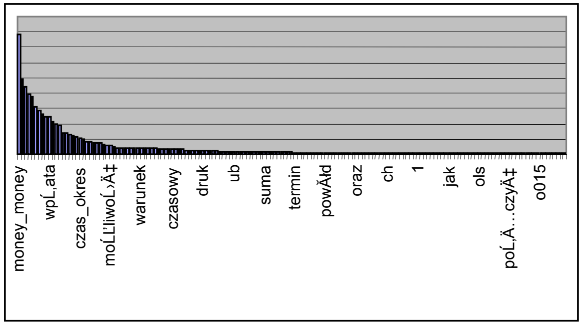
It *was a main building block of a platform for a hotline „customer remark database” analyser*

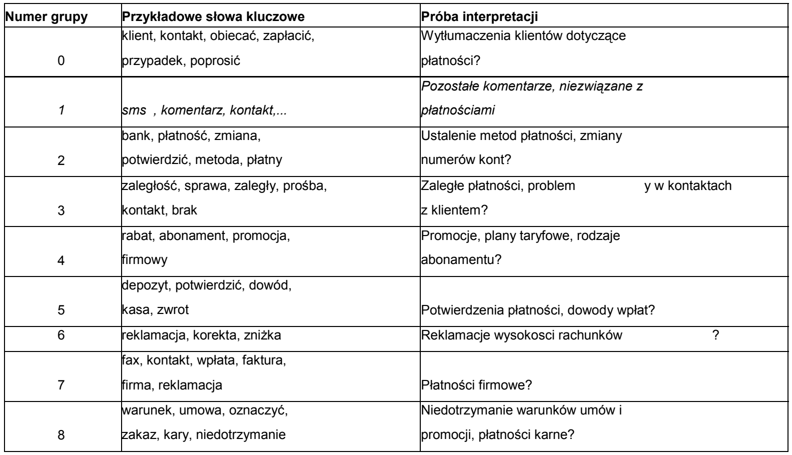
Main applications were:

- Remark database clustering in order to assess customer complaints type groups
- Classification of customers based on their compliants profile
- „Early warning” compliant detection in real time



The system will be used in a next project undertaken by WUT which starts in January 2005. This time the client is France Telecom, and XXX it will be used as a knowledge analysis engine in an experimental document management system.





Numer grupy	Przykładowe słowa kluczowe	Próba interpretacji
0	klient, kontakt, obiecać, zapłacić, przypadek, poprosić	Wytłumaczenia klientów dotyczące płatności?
1	*sms , komentarz, kontakt,...*	*Pozostałe komentarze, niezwiązane z płatnościami*
2	bank, płatność, zmiana, potwierdzić, metoda, płatny	Ustalenie metod płatności, zmiany numerów kont?
3	zaległość, sprawa, zaległy, prośba, kontakt, brak	Zaległe płatności, problem y w kontaktach z klientem?
4	rabat, abonament, promocja, firmowy	Promocje, plany taryfowe, rodzaje abonamentu?
5	depozyt, potwierdzić, dowód, kasa, zwrot	Potwierdzenia płatności, dowody wpłat?
6	reklamacja, korekta, zniżka	Reklamacje wysokosci rachunków ?
7	fax, kontakt, wpłata, faktura, firma, reklamacja	Płatności firmowe?
8	warunek, umowa, oznaczyć, zakaz, kary, niedotrzymanie	Niedotrzymanie warunków umów i promocji, płatności karne?

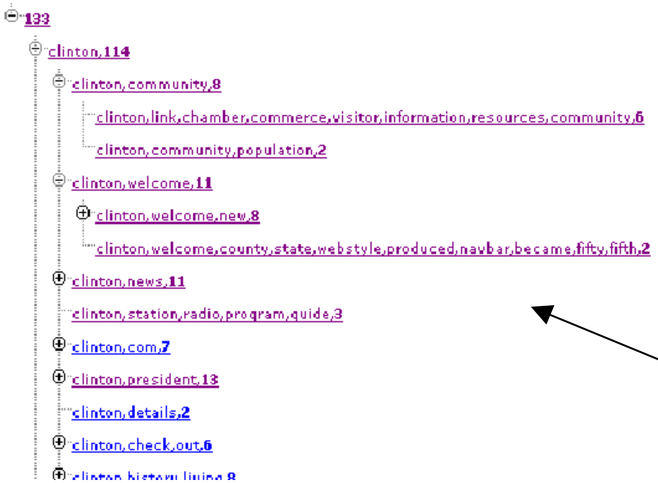
FAO database search improvements based on TM

FAO bibliographical databases based on UNESCO CDS/ISIS

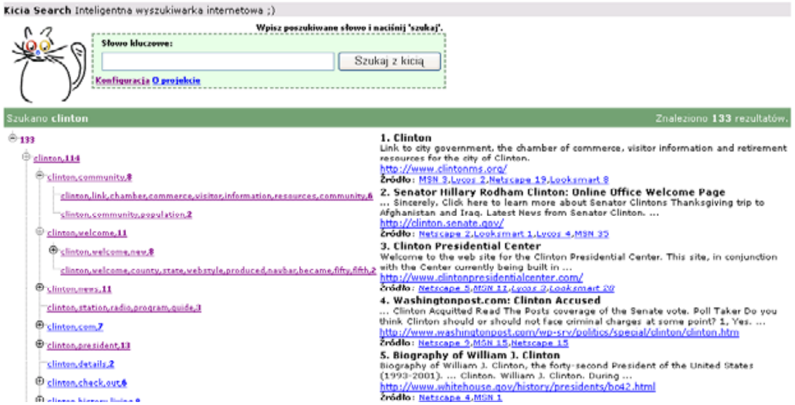
•***FAOBIB*** *(multilingual, on-line catalogue of documents and publications produced by FAO since 1945)*
•***AGRIS*** *(international information system for the agricultural sciences and technology, contains references to literature)*
•***CARIS*** *(current agricultural projects database, contains agricultural projects descriptions)*
•*other*

Web access provided by ICIE WWW/ISIS
Good quality standard search functions (field indexing, index lookup, cross-referencing)
Static thesaurus search assist (query building) is provided (via AGROVOC thesaurus)

The following TM based extensions are being developed (based on XXX system):
- TM assisted analysis of user query with keyword clustering & substitution
- automatic clustering of search results
- retrieved document set visualisation in a form of a graphical topic map



```
⊖133
  ⊖clinton,114
    ⊖clinton,community,8
      clinton,link,chamber,commerce,visitor,information,resources,community,6
      clinton,community,population,2
    ⊖clinton,welcome,11
      ⊕clinton,welcome,new,8
      clinton,welcome,county,state,webstyle,produced,navbar,became,fifty,fifth,2
    ⊕clinton,news,11
    clinton,station,radio,program,guide,3
    ⊕clinton,com,7
    ⊕clinton,president,13
    clinton,details,2
    ⊕clinton,check,out,6
    ⊕clinton,history,living,8
```



Kicia Search Inteligentna wyszukiwarka internetowa ;)

Wpisz poszukiwane słowo i naciśnij 'szukaj'.

Szukano **clinton** Znaleziono **133** rezultatów.

⊖133
⊖clinton,114
⊖clinton,community,8
clinton,link,chamber,commerce,visitor,information,resources,community,6
clinton,community,population,2
⊖clinton,welcome,11
⊕clinton,welcome,new,8
clinton,welcome,county,state,webstyle,produced,navbar,became,fifty,fifth,2
⊕clinton,news,11
clinton,station,radio,program,guide,3
⊕clinton,com,7
⊕clinton,president,13
clinton,details,2
⊕clinton,check,out,6
⊕clinton,history,living,8

1. Clinton
Link to city government, the chamber of commerce, visitor information and retirement resources for the city of Clinton.
http://www.clintonms.org/
Źródło: MSN 3,Lycos 2,Netscape 19,Looksmart 8

2. Senator Hillary Rodham Clinton: Online Office Welcome Page
... Sincerely. Click here to learn more about Senator Clintons Thanksgiving trip to Afghanistan and Iraq. Latest News from Senator Clinton. ...
http://clinton.senate.gov/
Źródło: Netscape 2,Looksmart 1,Lycos 4,MSN 35

3. Clinton Presidential Center
Welcome to the web site for the Clinton Presidential Center. This site, in conjunction with the Center currently being built in ...
http://www.clintonpresidentialcenter.com/
Źródło: Netscape 5,MSN 11,Lycos 3,Looksmart 28

4. Washingtonpost.com: Clinton Accused
... Clinton Acquitted Read The Posts coverage of the Senate vote. Poll Taker Do you think Clinton should or should not face criminal charges at some point? 1, Yes. ...
http://www.washingtonpost.com/wp-srv/politics/special/clinton/clinton.htm
Źródło: Netscape 9,MSN 15,Netscape 15

5. Biography of William J. Clinton
Biography of William J. Clinton, the forty-second President of the United States (1993-2001). ... Clinton. William J. Clinton. During ...
http://www.whitehouse.gov/history/presidents/bc42.html
Źródło: Netscape 4,MSN 1

System is being tested on web-based information sources with WWW/ISIS integration under way (through ICIE developed ISISDBC). Additionally a module for bibliographical notes noise management is being developed (automatic correction of bibliographical information typed in by librarians).

Mining in FAO databases

FAO databases are also a valuable source of information *per se* that could be used for automatic ontology building via discovering interesting semantic relationships between concepts and keywords

We started experiments with:

- **AGROVOC** network analysis using web graph analysis algorithms (for example hubs/authorities algorithms)
- Identification of important keyword groups
- Temporal analysis of **AGROVOC** changes

- Using Latent Semantic Analysis (SVD based coocurence analysis) and episodic rules to identify relationships between keywords in FAO full text database (**FAODOC**)

Conclusions

The resources mobilized for SEMKOS project have been not wasted New project within FP6 framework including old SEMKOS partners could – and should - be considered

Results:

- *XXX TM system has been developed*
- *Experiments with combining TM with bibliographical systems seem to be very promising*
- *Initial mining experiments in bibliographical databases have been started*

Thank you for your attention