Knowbel – EarlyTracks

S^2E – *Semantic Search Engine*

Knowbel
EarlyTracks
From Text To Knowledge

Knowbel – Earlytracks
○○

Our approach
○○○○

Technical functionalities
○○○○○○

Application to the media industry

Table of Content

Knowbel – Earlytracks

About us

Knowbel – Earlytracks combines the strength of two worlds :

- ► Knowbel Technologies, university Spin-Off
    - ► Origin: 10 years of research in Natural Language Processing
    - ► Main focus: medical & media

- ► Earlytracks, ventured back Start-Up
    - ► Origin: professional of both finance and entrepreneurship
    - ► Main focus: corporate finance & media

Mid-2012, Knowbel – Earlytracks decided to merge their activities. Since then, they moved to a fully commercial company.

Knowbel – Earlytracks
Core business

«**Text Mining** is the discovery by computer of new, previously
underlined:unknown information, by automatically extracting information
from different underlined:written resources. A key element is the linking
together of the extracted information together to underlined:form new facts
or new hypotheses to be explored further by more conventional
means of experimentation.» (Hearst, 2003)

- ▶ **Three main domains:** medical, media & corporate finance
- ▶ **Two main applications:**
    - ▶ Pull: the request for a given transaction is initiated by the receiver
      (*i.e. search engine*)
    - ▶ Push: the request for a given transaction is initiated by the
      publisher (*i.e. survey engine*)

Our approach

The goal behind

Our approach
Why not an ontology?

An ontology is constrained by:

- ▶ complexity of the domain
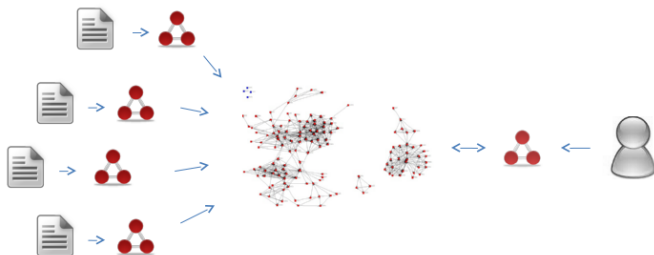- ▶ stablity of the domain
- ▶ homogeneity of the domain

→ Solution: divide into sub-domain?

Not really... risk of semantic heterogeneity!

- ▶ logical interpretations
- ▶ structural representation
- ▶ naming and format specification

Knowbel – Earlytracks
○○

Our approach
○○●○

Technical functionalities
○○○○○○

Application to the media industry

Our approach
A bottom-up design

Our approach
Two-level indexation

Go beyond the words

- ▶ Text Level
  - ▶ Multiword units: several words, one meaning

    EN: chief executive officer, net income, ...
    FR: chambre des représentants, commission d'enquête parlementaire, ...

- ▶ Semantic Level
  - ▶ Named-Entities: predefined domain categories

    Health Care: drugs, diseases, vital signs, ...
    Finance: companies, persons, market index, ...

  - ▶ Events: set of relationships linking one or more entities

    Biography: weddings, professions, death, ...
    Finance: bankruptcy, joint ventures, acquisition, ...

Technical functionalities

Content enrichment (1)

Technical functionalities

Content enrichment (2)

Technical functionalities

Content structuration (1)

Technical functionalities

Content structuration (2)

Technical functionalities

Content contextualization (1)

Technical functionalities

Content contextualization (2)

Application to the media industry

- ▶ **Editorialism**
    - ▶ concurrence survey (news not covered, ...)
    - ▶ identification of plagiarism
    - ▶ contextualization (pictures, videos, archives, ...)
    - ▶ ....

- ▶ **User side**
    - ▶ transparent query expansion
    - ▶ vertical search (facet browsing, ...)
    - ▶ profiling (dynamic portal, ...)
    - ▶ ...

$\rightarrow$ Main interests are:

- ▶ real-time update
- ▶ certified information