Mindex as a Source of Topic Map

Sachio Hirokawa

Project Lafla
Research Institute for Information Technology
Kyushu University

@AToMS - Asian Topic Maps Summit 2007, Kyoto

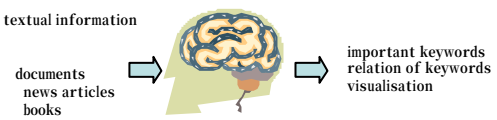
outline

- Human Information Processing
  - input,processing&discussion,output
- Concept Graph
  - The meaning of documents is determined by the words in the documents. The meaning of words is determined by the documents that contain the words.
- Examples of Concept Graph
  - Newspaper, Scientific Journals,Patents,Dictionary, Questionnaire
- Mindex—Next Generation Search Engine

INPUT

ANALYSE & THINK

DISCUSS & DISPLAY



Concept Graph

- The meaning of documents is determined by the words in the documents. The meaning of words is determined by the documents that contain the words.
- Concept Graph search engine extracts Characteritic Words and visualises their hypenym/supernym relationship.

Characteristic Words

- A word has different meaning according to the context where the word is used. The semantical relationship of two words depends also on the context where they are used. We formalize the context simply as a set of documents and formalize hypernym/hyponym relation according to the document frequencies of words.
- Assume that U shows the set of whole documents. Given a query q, D(q) represent the set of documents that satisfy the query q. Given a word w, and a set of documents X, df(w;X) represents the number of documents in X that contains the word w. A word w is characteristic if df(w;D(q))=df(w; U) > 0.5. In other words, a word is characteristic to the search result D(q) when more than half documents that contains the word w belong to D(q). If we say more roughly, the word w is characteristic to the query q, when almost all documents that contain the word w satisfy the query q.

Hypernym/Hyponym Relation

- We introduced a formulation of hypernym/hyponym relation of words according to the document frequencies. A word u is a hypernym of v with respect to D(q) when they satisfy the following two conditions.
  df(u; U) > df(v; U)
  df(u v;D(q))=df(v;D(q)) > 0.5
- Here, df(u v;D(q)) represents the number of documents in D(q) that contain both u and v. In other words, u is a hypernym of v, when u occurs more often in all documents than v and most documents that contain v contains u.

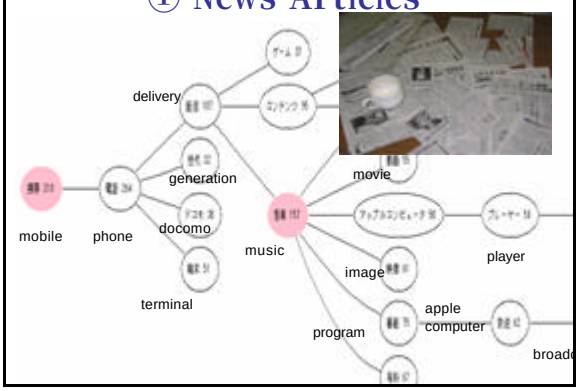
Direct Upper Relation

- Hypernym/hyponym relation determines an order structure among characteristic words and can be drawn as a directed acyclic graph. However, some words have too many hypernym and the graph may contain edges overlapped each other. To obtain more clear structure, we define "direct upper/lower" relation between words.
- Given a word v, the set UP(v) of upper words of v and the set DUP(v) of direct upper words of v are defined as follows.

  UP(v) = { u | D(q) is a hypernym of v }
  DUP(v) = { u | ∀w ∈ DUP(v) ¬(v ∈ UP(w)) }

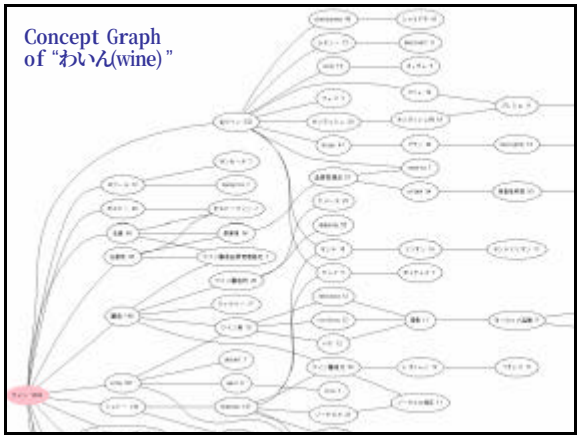
- A hypernym u of v is a direct upper hypernym when there are no other hypernym of v between u and v. Visualization of a concept graph can be obtained by placing words of high frequencies on the left sides and ones with lower frequencies on the right sides.

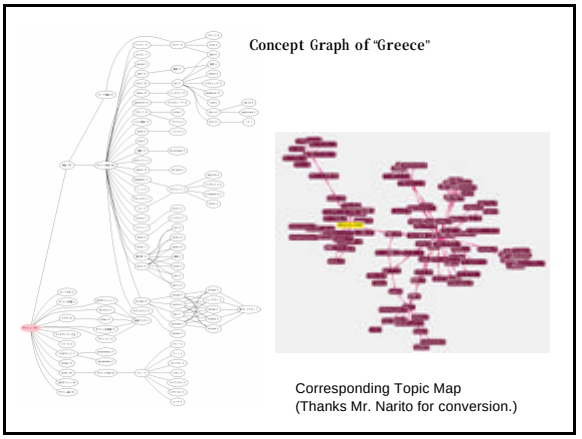
① News Articles

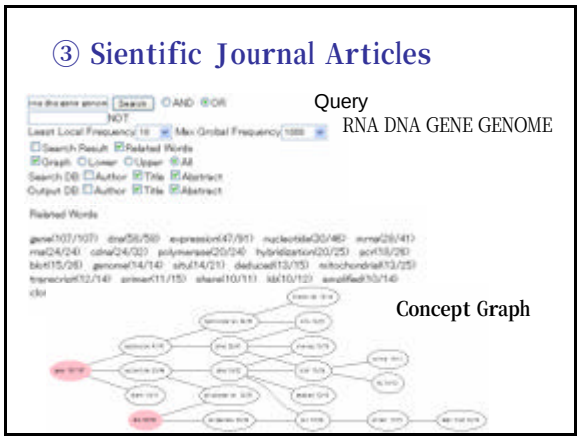


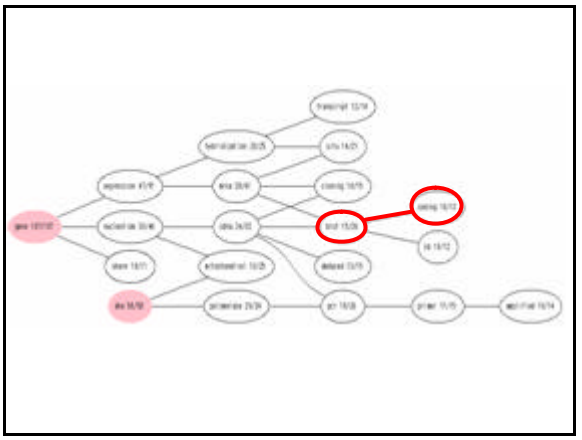
② English - Japanese Dictionary

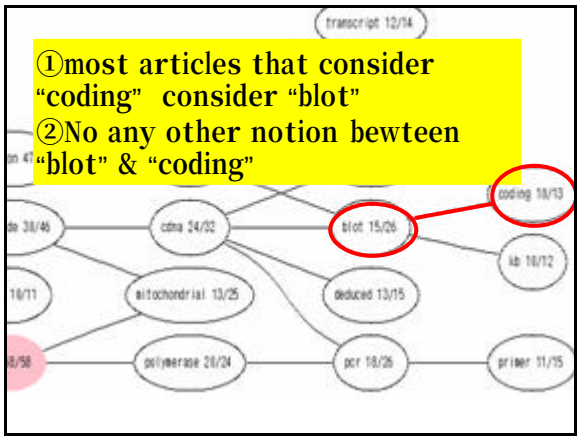
document unit	description of a word
#documents	1,648,628
#words	986,410
#occurrences of words	8,644,997
total size(bytes)	112,624,437
average size of documents(byte)	68.3

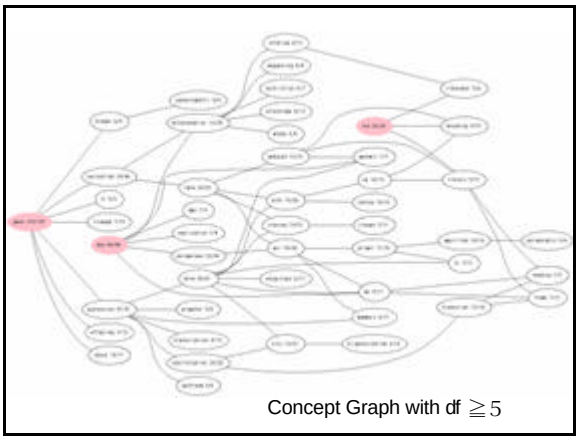

Concept Graph
of “わいん(wine)”


Concept Graph of “Greece”
Corresponding Topic Map
(Thanks Mr. Narito for conversion.)


③ Sientific Journal Articles
Query
RNA DNA GENE GENOME
Concept Graph



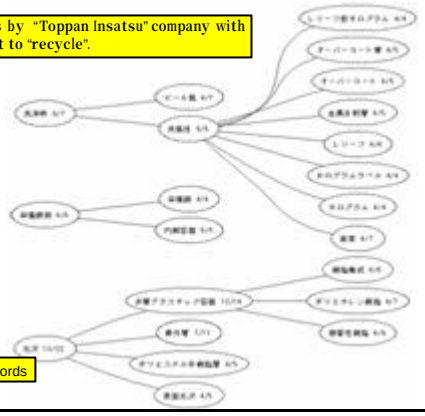

①most articles that consider “coding” consider “blot”
②No any other notion bewteen “blot” & “coding”
transcript 12/14
coding 18/13
cdna 24/32
blot 15/26
kb 18/12
mitochondrial 13/25
deduced 13/15
polymerase 20/24
pcr 18/26
primer 11/15


Concept Graph with df ≧5

④ Patent Documents

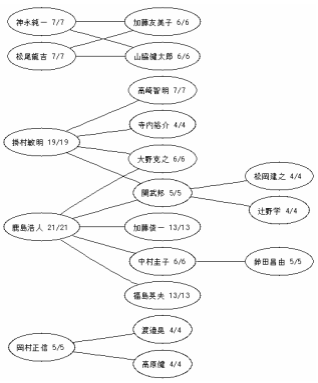
From:National Center for Industrial Property Information and Training
http://www.ipdl.inpit.go.jp/homepg.ipdl

Pattens by "Toppan Insatsu" company with respect to "recycle".

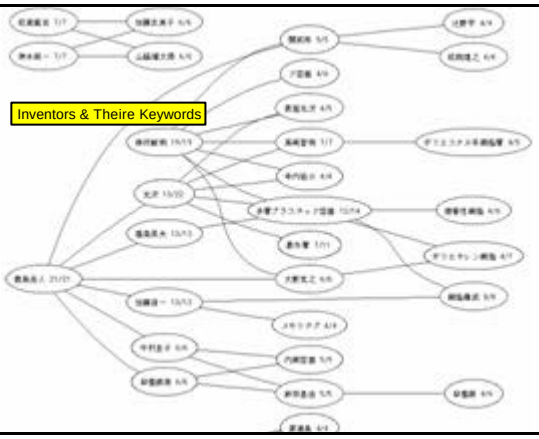


Keywords

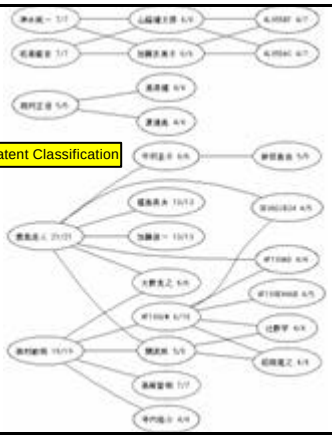
Inventors



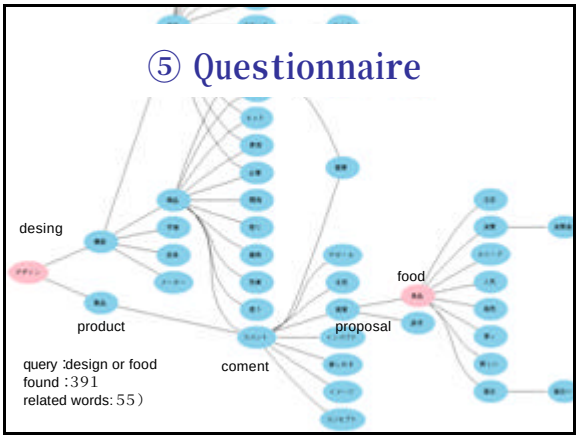
Inventors & Theire Keywords



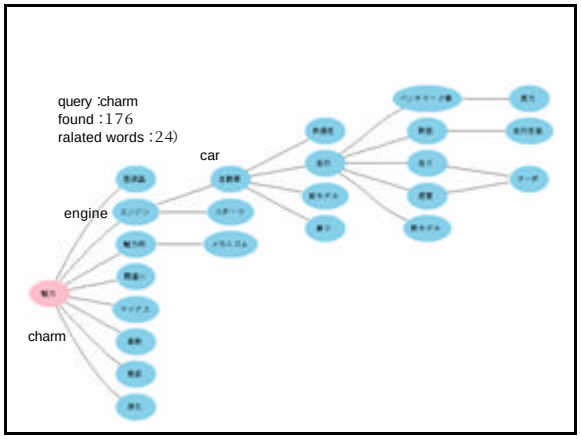
Inventors & Patent Classification

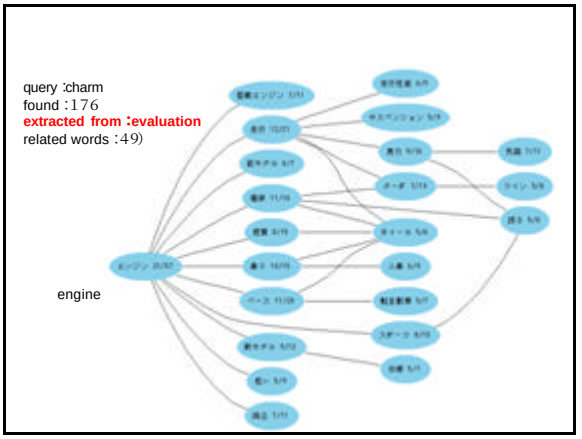


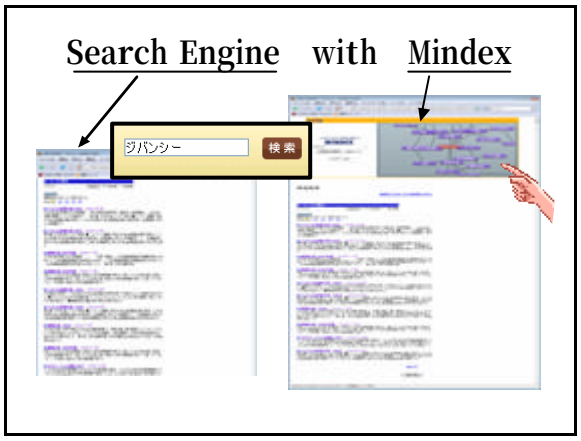
⑤ Questionnaire

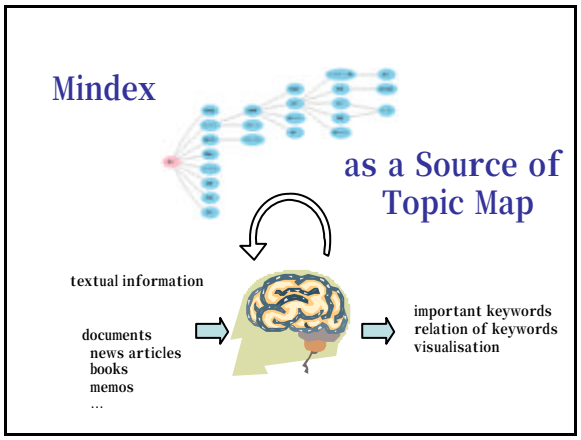


query :design or food
found :391
related words: 55)


query :charm
found :176
ralated words :24)
car
engine
charm


query :charm
found :176
extracted from :evaluation
related words :49)
engine


Search Engine with Mindex
ジバンシー
検索


Mindex
as a Source of
Topic Map
textual information
documents
news articles
books
memos
...
important keywords
relation of keywords
visualisation