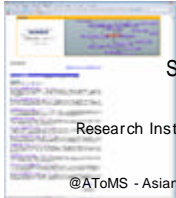


Mindex as a Source of Topic Map



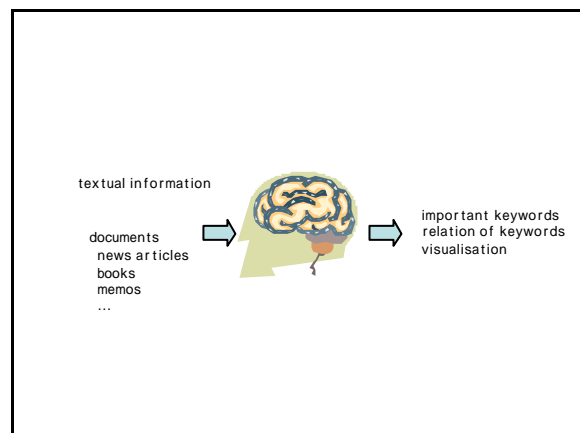
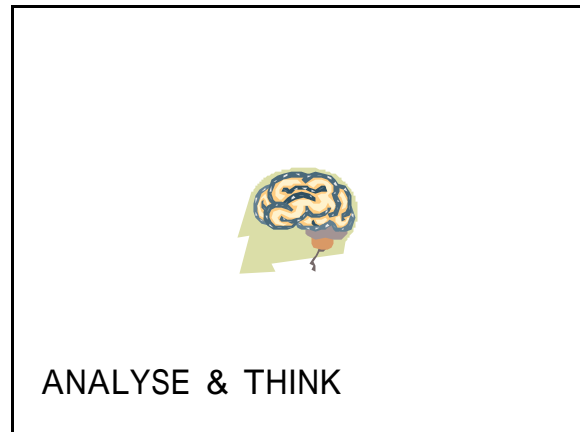
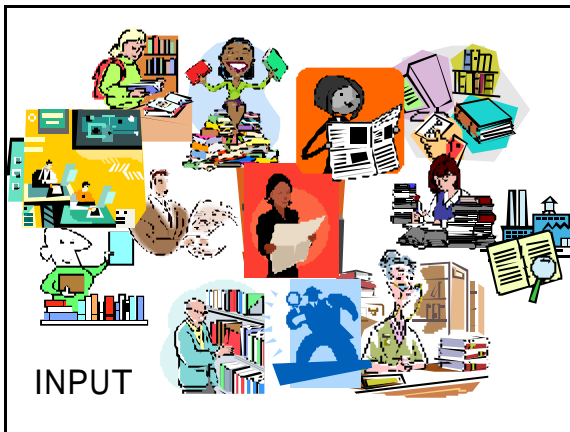
Sachio Hirokawa

Project Lafia
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



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 Characteristic 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 hypenym/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 hypenym/hyponym relation of words according to the document frequencies. A word u is a hypenym 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 hypenym of v , when u occurs more often in all documents than v and most documents that contain v contains u .

Direct Upper Relation

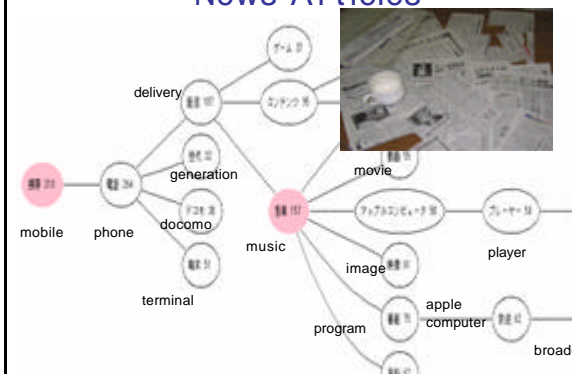
- ≒ Hypernym/hyponym relation determines an order structure among characteristic words and can be drawn as a directed acyclic graph. However, some words have too many hypenym and the graph may contain edges overlapped each other. To obtain more clear structure, we define "direct upper / lower" relation 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.

$$UP(v) = \{u \mid D(q) \text{ is a hypenym of } v\}$$

$$DUP(v) = \{u \mid w \text{ } DUP(v) \sim (v \text{ } UP(w))\}$$

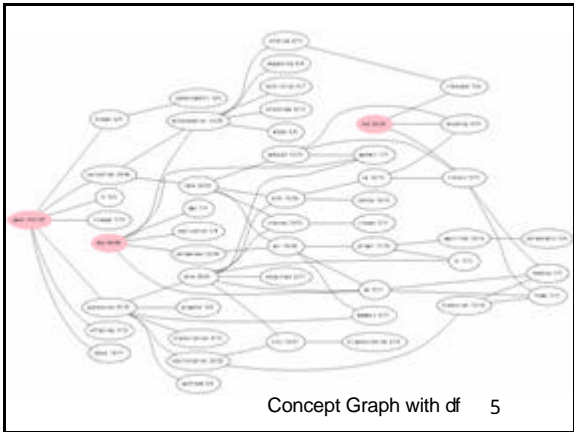
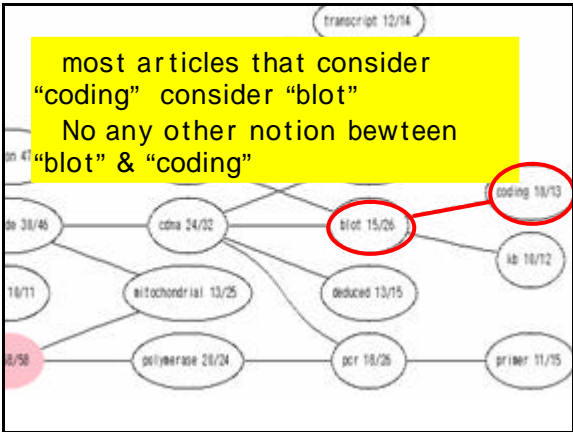
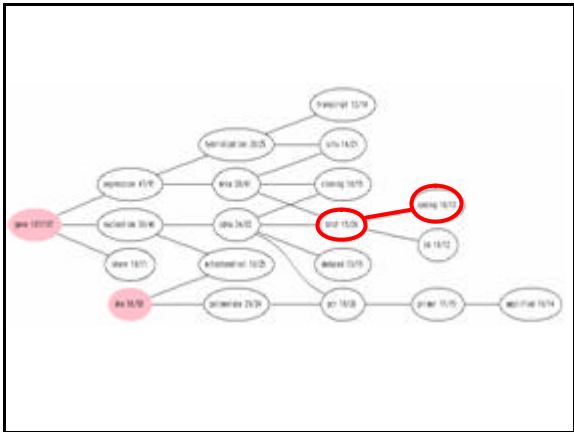
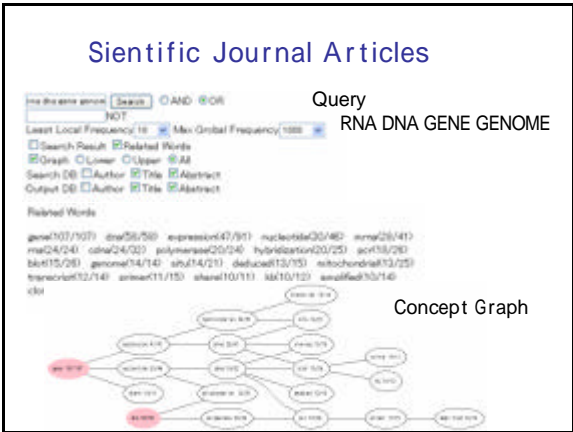
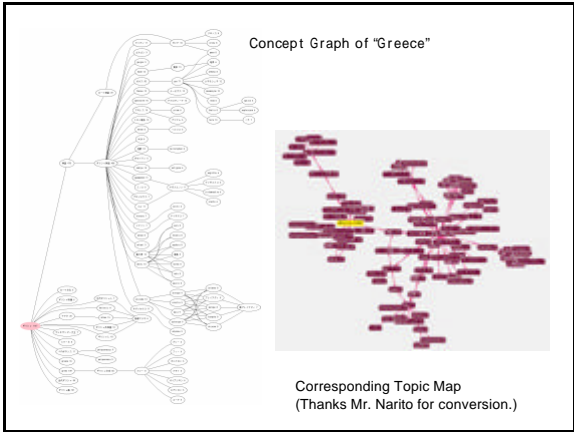
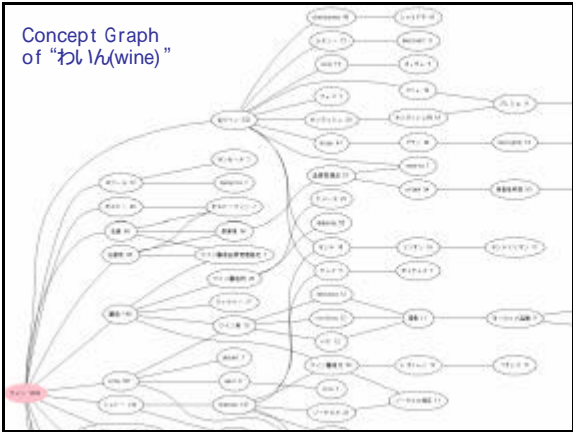
- ≒ A hypenym u of v is a direct upper hypenym when there are no other hypenym of v between u and v . Visualization of a concept graph can be obtained by placing words of high frequencies on the left sides and ones with lower frequencies on the right sides.

News Articles



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

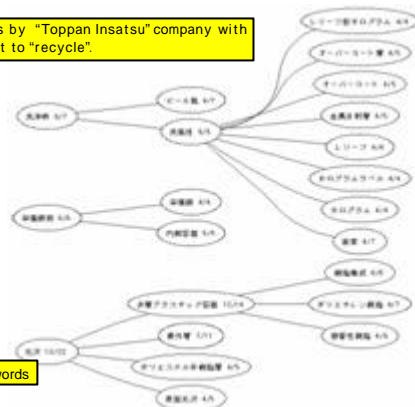


Patent Documents



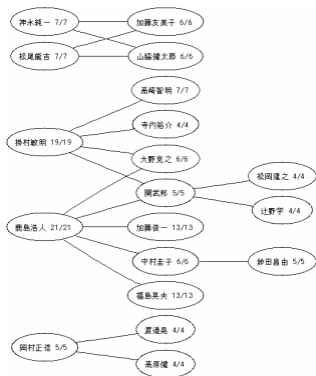
From: National Center for Industrial Property Information and Training
<http://www.ipdl.inpit.go.jp/homepg.ipdl>

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

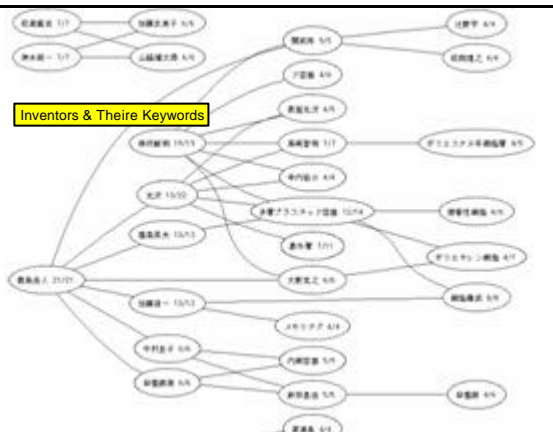


Keywords

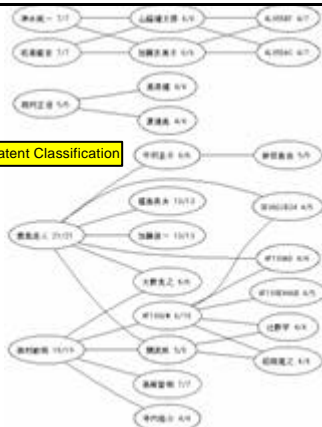
Inventors



Inventors & Their Keywords



Inventors & Patent Classification



Questionnaire

