How do We Incorporate a Recommendation Framework into the Search Engines?

November 17th, 2008 (Kobe, Japan)

Tatsumi Kobayashi

Yahoo! JAPAN

YAHOO! JAPAN

Today's Agenda

1. *Challenge for Next Generation Search*

   Towards 4th Generation Search

2. *Analysis and Discovery*

   Collective Intelligence for Topic and Trend Analysis

3. *Recommendation Strategy*

   Model-based approach for adaptation to user behaviors and dynamic lexical sense change

4. *Conclusion and Future Work*

Chapter 1

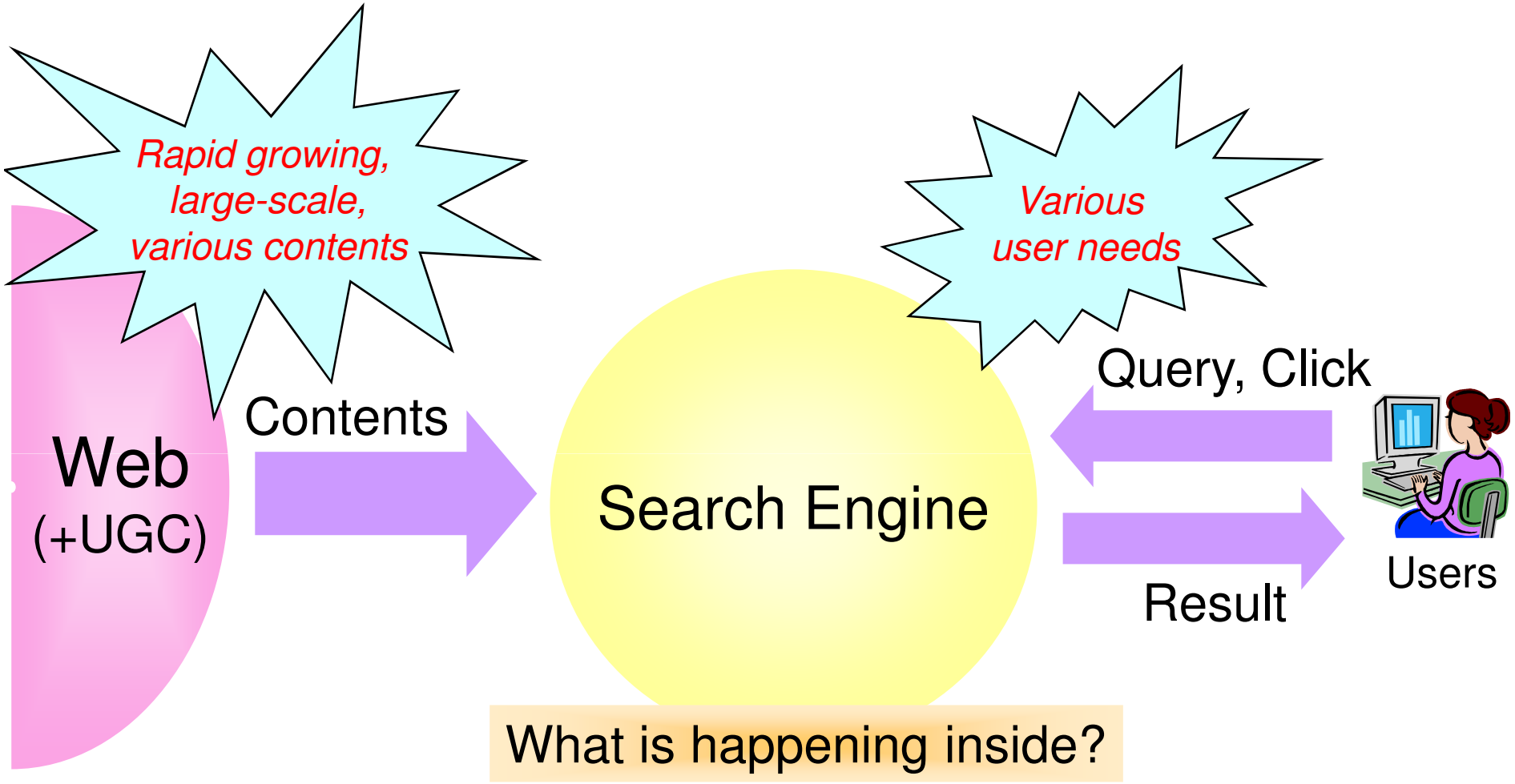
Challenge for Next Generation Search

Towards 4th Generation Search

***Yahoo! JAPAN** is a market leader of web search engines in JAPAN !!*

YAHOO! JAPAN

Recent Changes on the Web



* UGC: User Generated Content (blog, SNS, etc.)

How are the Search Engines Struggling?

Query side

[Problem]

Very frequent change of query meaning

[Current strategy]

Spelling suggestion, query rewriting, query suggestion, etc.

Content side

[Problem]

Complexity and its very fast change of contents

[Current strategy]

To blend UGC and news to web search results independently

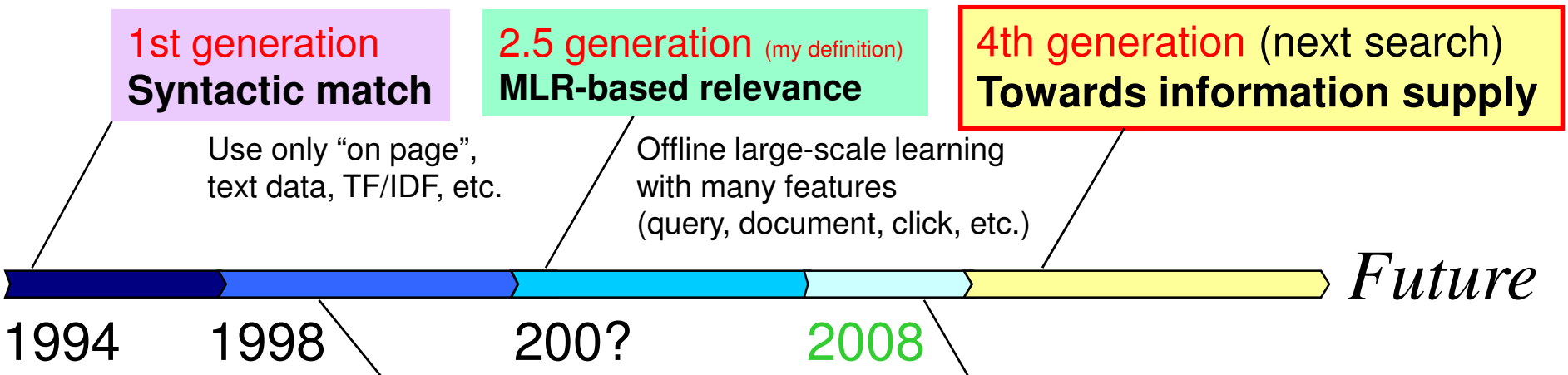
To try to crawl the web as fast as possible

It's not enough!!!

The Evolution of Commercial Web Search Engines

(Broder's definition [ECIR 2007])

What is this?



Help the users

Spell checking, query suggestion/rewrite, context help (assisted search), etc.

Integrates multiple sources of data

Blending of UGC, news, house-made contents to the results

* MLR: Machine Learning Ranking

We are now in the 3rd generation search!

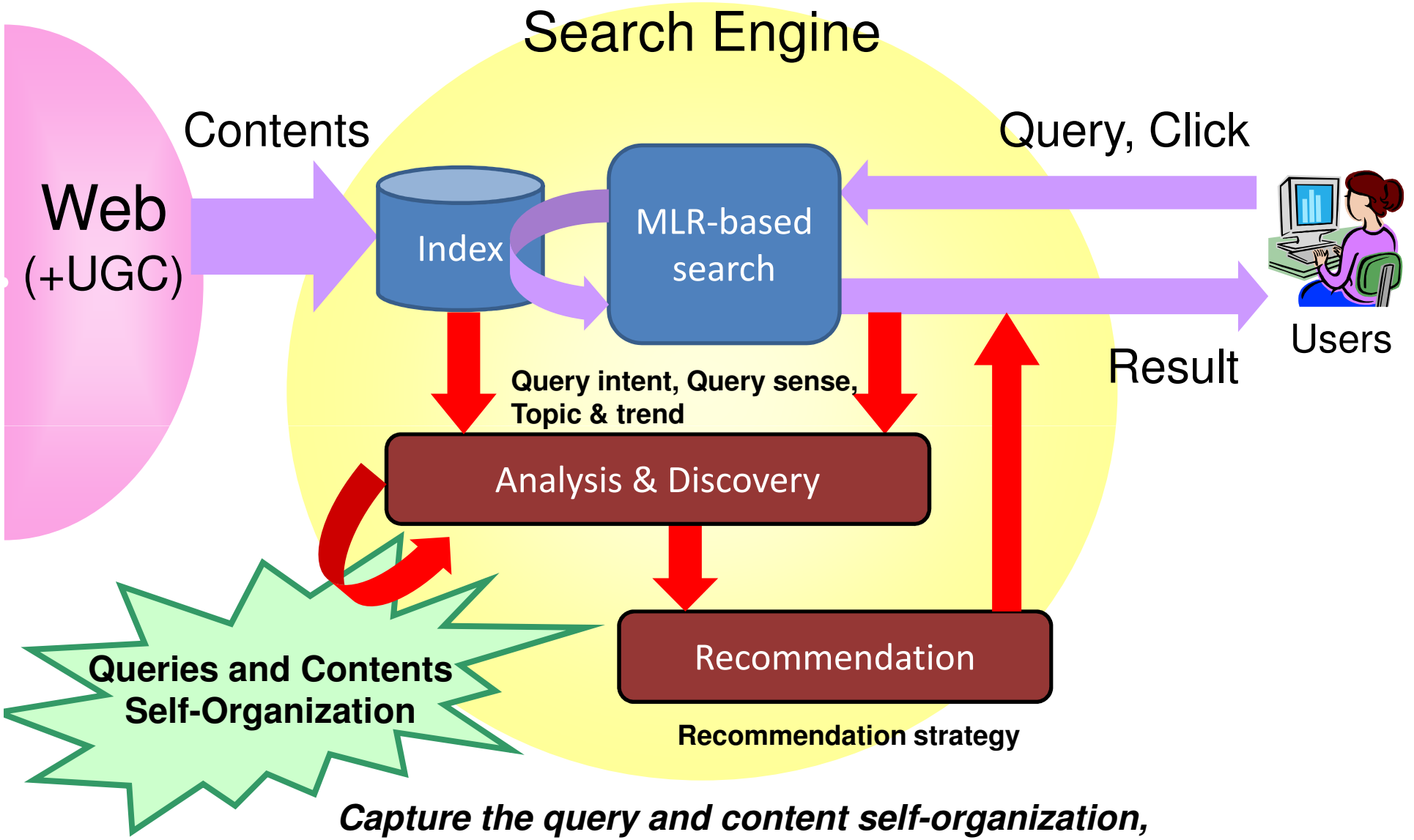
YAHOO! JAPAN

Crucial Issues from 3rd to 4th Generation Search

1. **Weakness of MLR for trend sensitivity**
   - Need another approach in addition to relevance calculation
2. **Need handling query intent and query sense**
   - Query intent analysis (3rd generation is going to handle)
   - Polysemy and similarity of lexical sense
   - Change and emergence of lexical sense
3. **How to discover new topics and trends**
   - To answer user's needs more accurately
4. **What is information supply?**
   - How to pick up information?
   - How to supply ("*recommend*") information?

We need a new recommendation framework on the top of the existing search engines

A Recommendation Framework Overview (Just one idea)



Capture the query and content self-organization, then create the smart recommendation system!

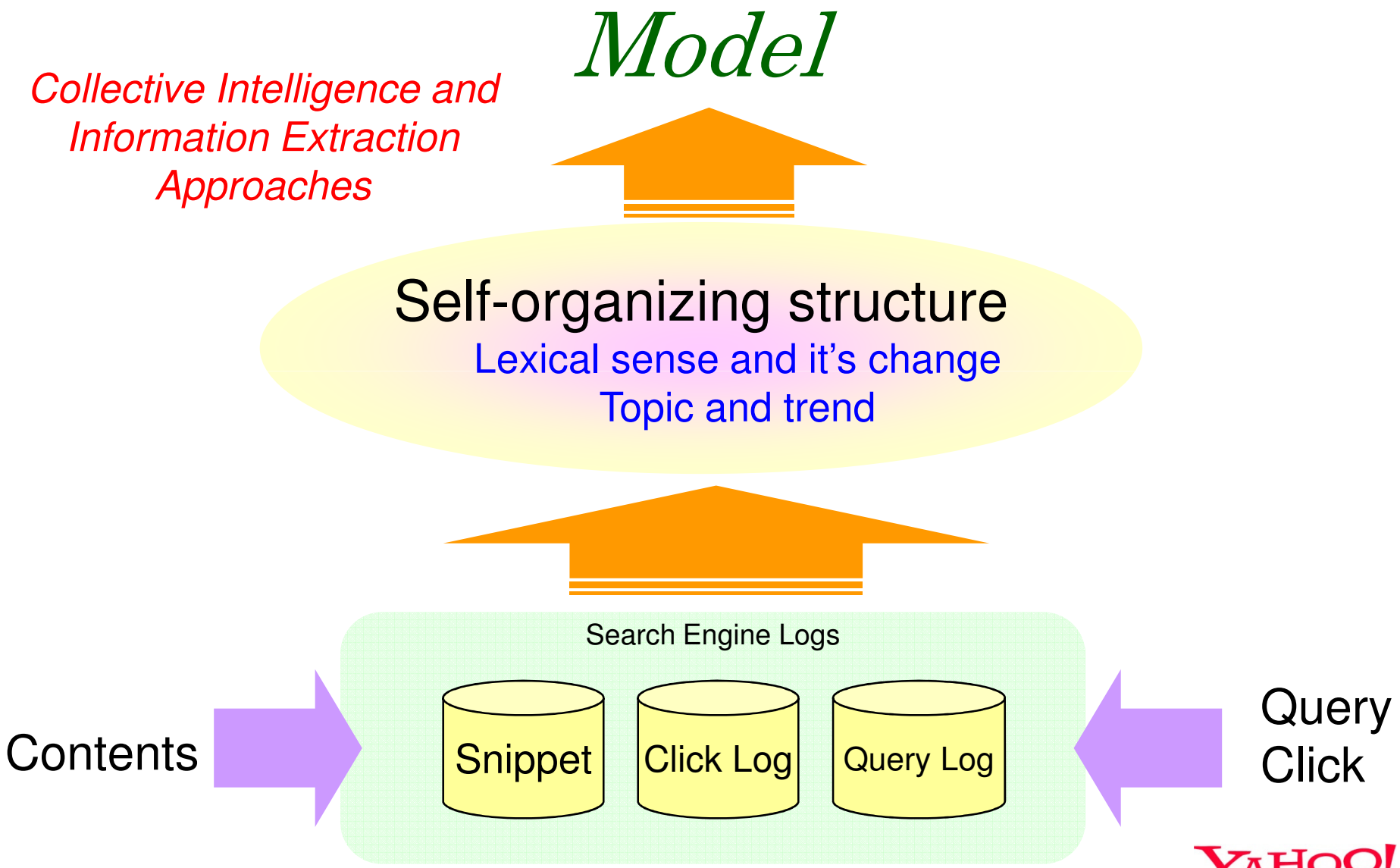
YAHOO! JAPAN

Chapter 2

Analysis & Discovery

Collective Intelligence
for Topic and Trend Analysis

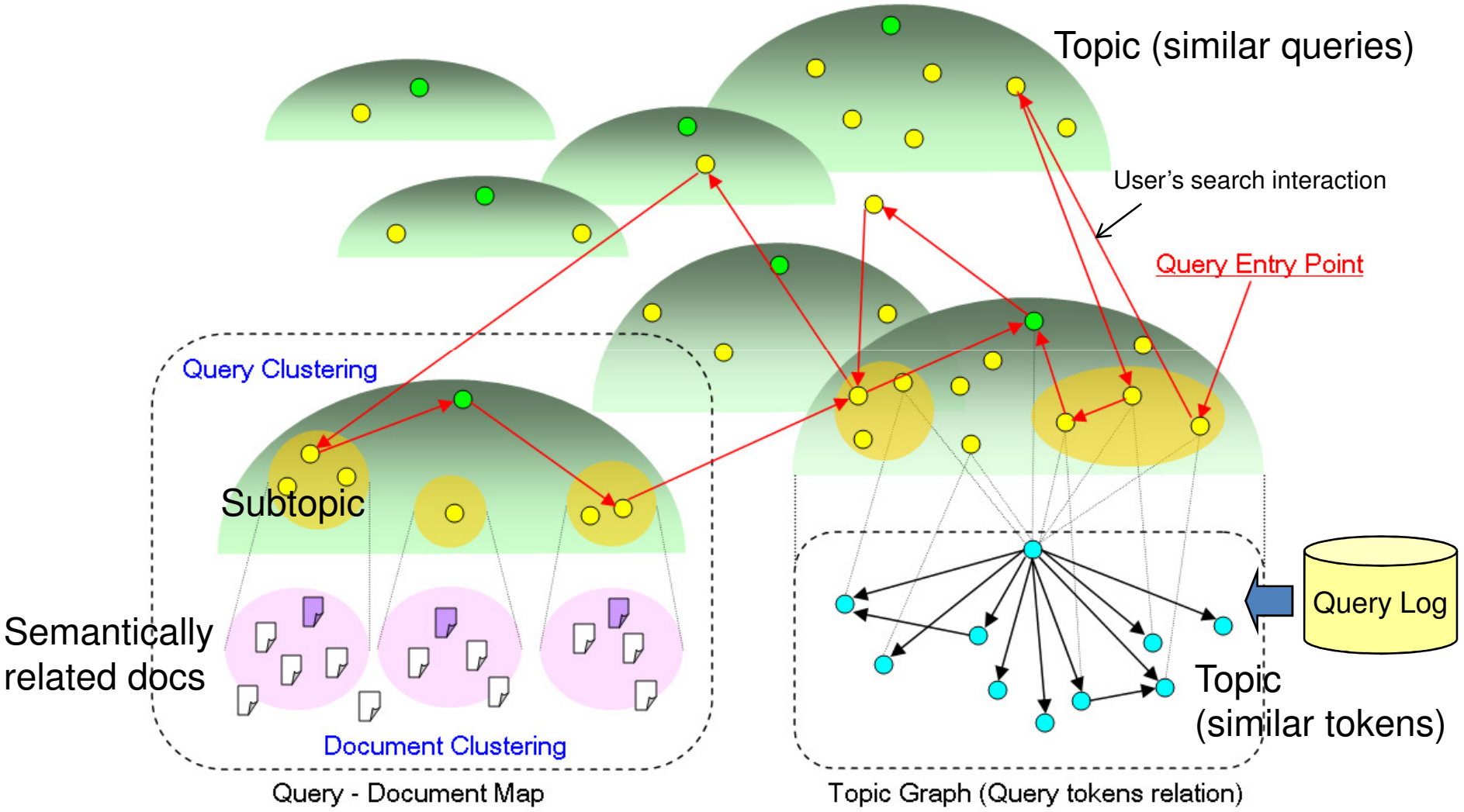
Goal of Analysis and Discovery Process



YAHOO! JAPAN

Self-Organization of User Interaction (Final Goal)

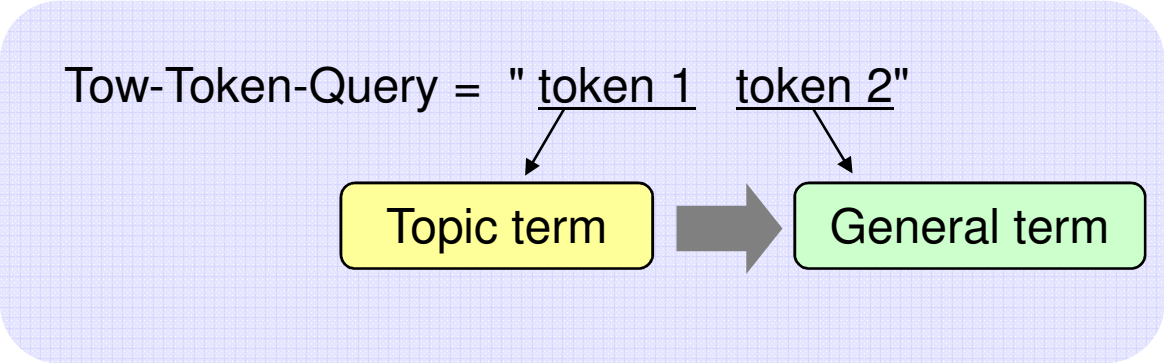
It's a structure emerged from user's mass behaviors.
We can map a user query on the structure.



YAHOO! JAPAN

Introducing "*Topic & General Token Pair*" Hypothesis

- Two-Token query has a general tendency statistically to form a "*topic token + general token*" pair

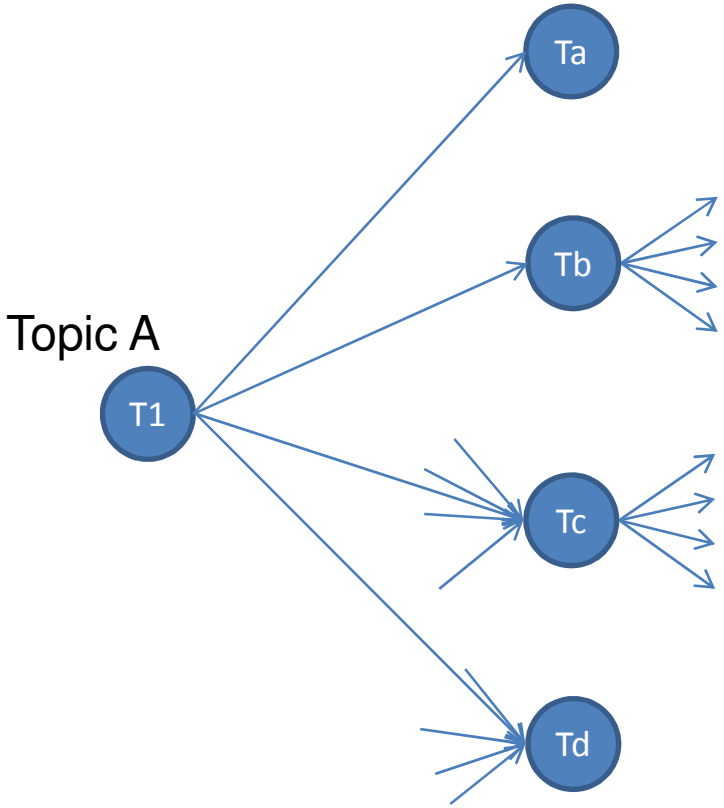


- Ex. "Olympic schedule"

- It is a strong property of CJK (Asian) Language due to white space delimiter.
  (CJK = Chinese 中国語, Japanese 日本語, Korean 韓国語)
  But the idea could be applied to non-CJK (English, etc.).

YAHOO! JAPAN

Properties of Two-Token Query in Graph



Ta has a strong relation to the topic A.

Tb has a strong relation to the topic A.
But it has an own topic property.

Tc has a weak relation to the topic A.
But it has an own topic property.

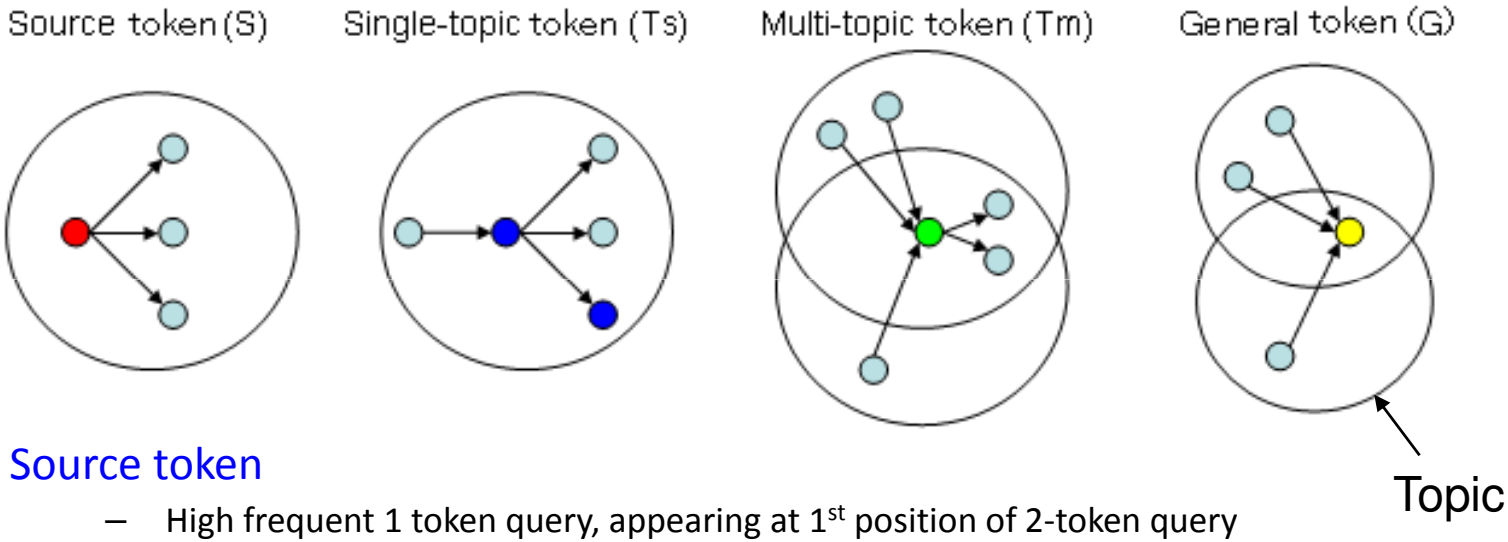
Td has a less relation to the topic A.
It has a strong general property

Look at number of parent nodes and children nodes

YAHOO! JAPAN

Topic Graph and Definition of Four Types of Tokens

- Based on such a two-token query idea, we can construct Topic Graph
- Four types of tokens are distinguished using # of link and frequency



Source token

– High frequent 1 token query, appearing at 1st position of 2-token query

Single-topic token

– Topic token appearing at 2nd position linking to one topic token at 1st position

Multi-topic token

– Topic token appearing at 2nd position linking to multiple topic tokens at 1st position

General token

– Few 1 token query, appearing frequently at 2nd position of 2-token query

YAHOO! JAPAN

Making Topic Graph: Algorithm 1/3

Step 1: Topic & general property calculation

Token strength (= Entropy)

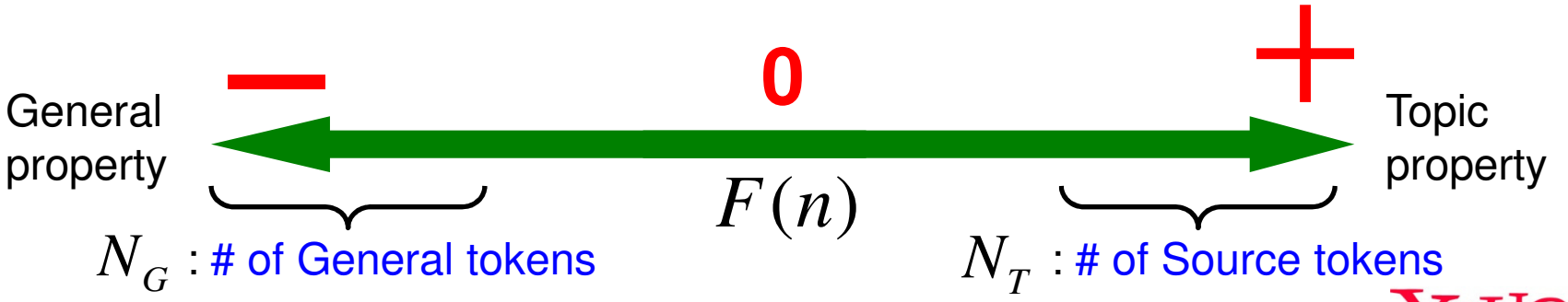
$$E(n) = -\frac{N_S(n) + N_1(n) + N_2(n)}{N_A} \log_2 \left(\frac{N_S(n) + N_1(n) + N_2(n)}{N_A} \right)$$

Topic/general degree

$$F(n) = \frac{\cancel{N_1(n) + N_2(n)}}{N_S(n) + N_1(n) + N_2(n)} \frac{N_1(n) - N_2(n)}{\cancel{N_1(n) + N_2(n)}} E(n)$$

2-token ratio — 1 or 2 pos token strength

N_S : # of 1-token query n (more than frequency of term N_U)
$N_1(n)$: # of 1st position token n in two-token query
$N_2(n)$: # of 2nd position token n in two-token query
N_A : # of all queries (more than frequency of term N_L)



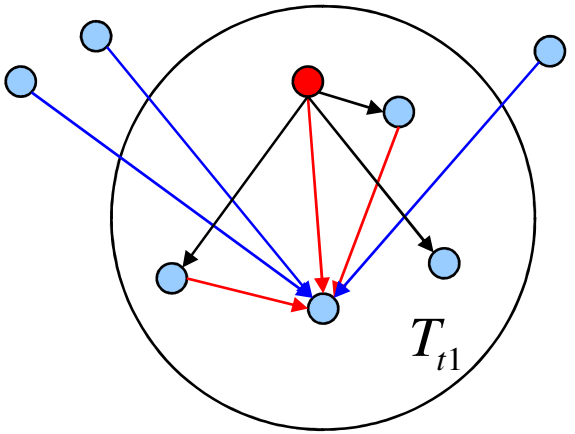
Making Topic Graph: Algorithm 2/3

Step 2: Topic clustering

For each source token, collect all 2nd position tokens linked from source token t1 in (t1 → t2), then put them to set T_{t1}

PMI (Pointwise Mutual Information)

$$PMI(t_1,t_2) = \log_2 \left\{ \frac{N(t_1,t_2)}{N(t_1,*)N(*,t_2)} \right\}$$



Ratio of PMI (all tokens vs. tokens in topic)

$$RPMI(t_1,t_2) = \frac{\sum_{i \in N_A} PMI(i,t_2)}{N_A} \frac{|T_{t1}|}{\sum_{i \in T_{t1}} PMI(i,t_2)}$$

Identify top N_R of high RPMI tokens in T_{t1} (Topic cluster)

Topic cluster set $T = \{T_i \mid i = \text{source topic tokens}\}$

Topic size $D(t) = \sum_{t \in T_i} E(t)$ (sum of topic strength of all tokens)

YAHOO! JAPAN

Making Topic Graph: Algorithm 3/3

Step 3: Synonym discovery using distributional similarity

Similarity coefficient (ordered)

$$Sim(t_1 \rightarrow t_2) = \frac{1}{2}\left\{\frac{N_b(t_1 \mid t_1 \in T_b(t_1) \cap T_b(t_2))}{N_b(t_1 \mid t_1 \in T_b(t_1))} + \frac{N_f(t_1 \mid t_1 \in T_f(t_1) \cap T_f(t_2))}{N_f(t_1 \mid t_1 \in T_f(t_1))}\right\}$$

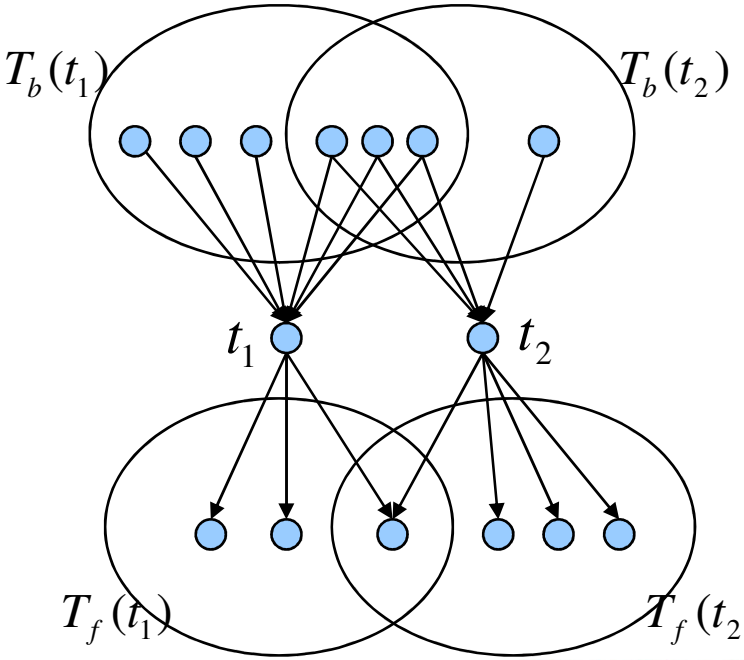
$$Sim(t_2 \rightarrow t_1) = \frac{1}{2}\left\{\frac{N_b(t_2 \mid t_2 \in T_b(t_1) \cap T_b(t_2))}{N_b(t_2 \mid t_2 \in T_b(t_2))} + \frac{N_f(t_2 \mid t_2 \in T_f(t_1) \cap T_f(t_2))}{N_f(t_2 \mid t_2 \in T_f(t_2))}\right\}$$

$N_b(t)$: # of source tokens to token t

$T_b(t)$: Source tokens set to token t

$N_f(t)$: # of target tokens from token t

$T_f(t)$: Source tokens set from token t



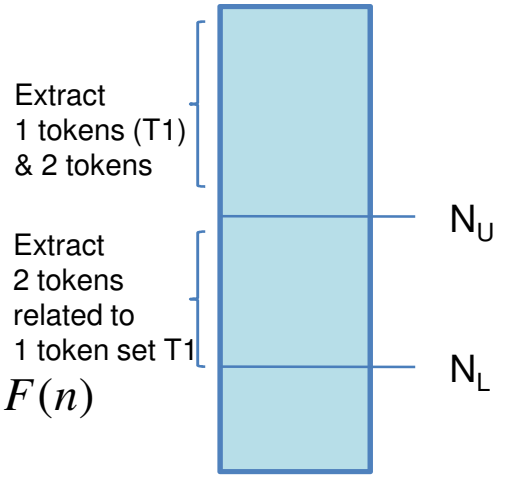
Find out contextually similar tokens

ex. {car, automobile}

YAHOO! JAPAN

Evaluation

- Data and conditions
  - One day query log (Aug 19th, 2008)
  - Frequency threshold $N_U = 5000, N_L = 10$



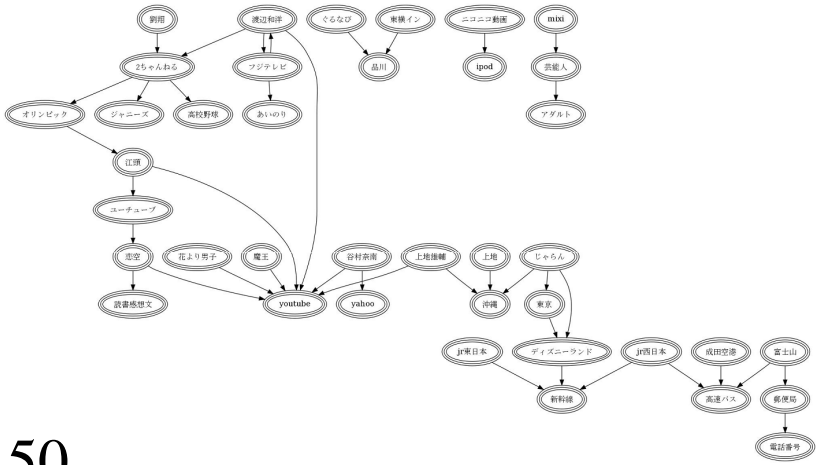
- Discovered tokens
  - Out of 100 source (topic) tokens in the top of the list $F(n)$
    - 30 trend tokens are found
      “北京オリンピック” (Beijing Olympic), “オリンピック” (Olympic),
      Various summer events (travel, high-school baseball, etc), TV, Movie, etc.
    - Others are major tokens: “Yahoo”, “Google”, “Toyota”, etc.
  - Out of 100 general tokens in the bottom of the list $F(n)$
    - “レシピ” (recipe), “動画” (moving image), “映画” (movie), “画像” (image), “ブログ” (blog), “地図” (map), “ゲーム” (game), “天気” (weather), “価格” (price), “wiki”, “無料動画” (free moving image), “辞書” (dictionary), etc.

YAHOO! JAPAN

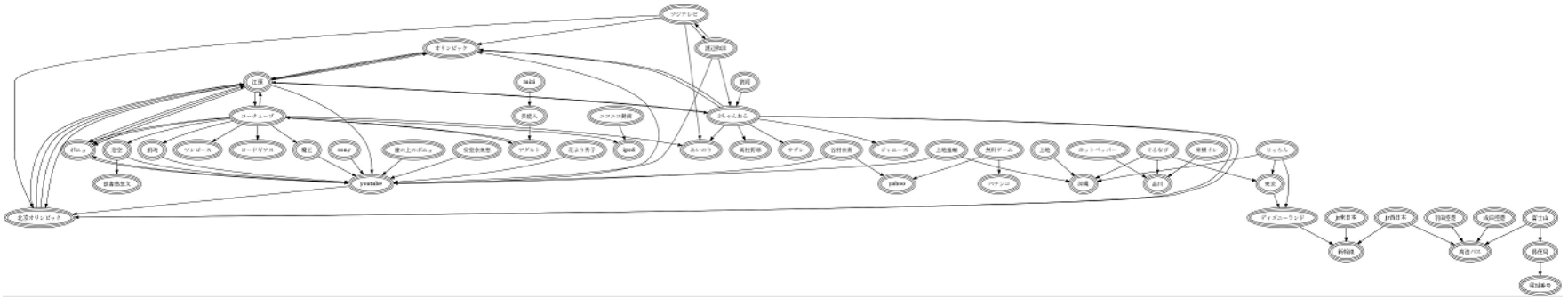
Top 100 Source Tokens Relations (except isolated tokens)

Source tokens have some of relations each other, sharing 2nd tokens

# of source token $N_T = 100,$ # of 2nd pos of source token $N_R = 30$



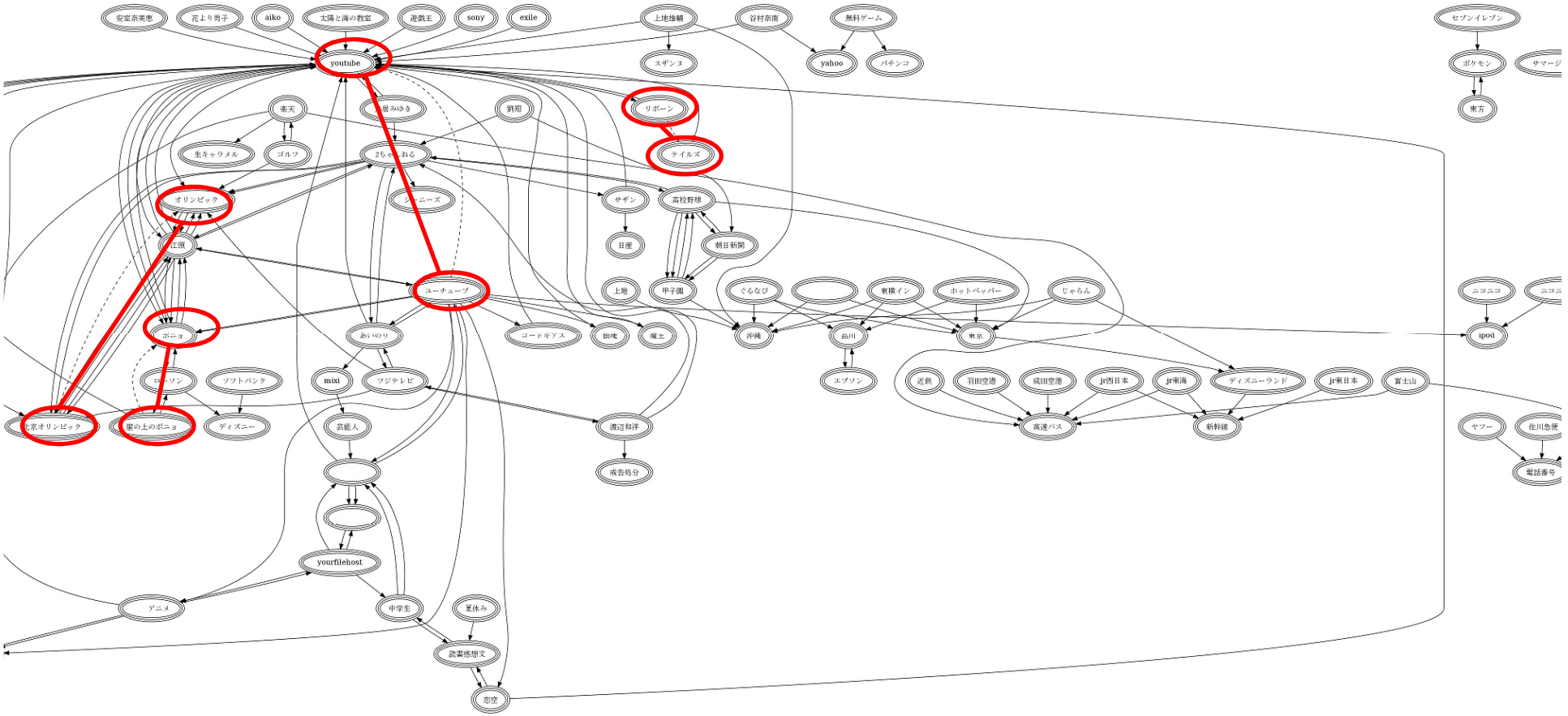
$N_T = 100, N_R = 50$



YAHOO! JAPAN

Similar Token Discovery (1/2)

$$N_T = 200, N_R = 30$$



{youtube, ユーチューブ} = different expression of alphabet and Katakana (Japanese)
{オリンピック, 北京オリンピック} = Olympic and Beijing Olympic
{崖の上のポニョ, ポニョ} = Long and short titles of Miyazaki Animation
{リボーン, テールズ} = Both two tokens are comics characters

YAHOO! JAPAN

Similar Token Discovery (2/2)

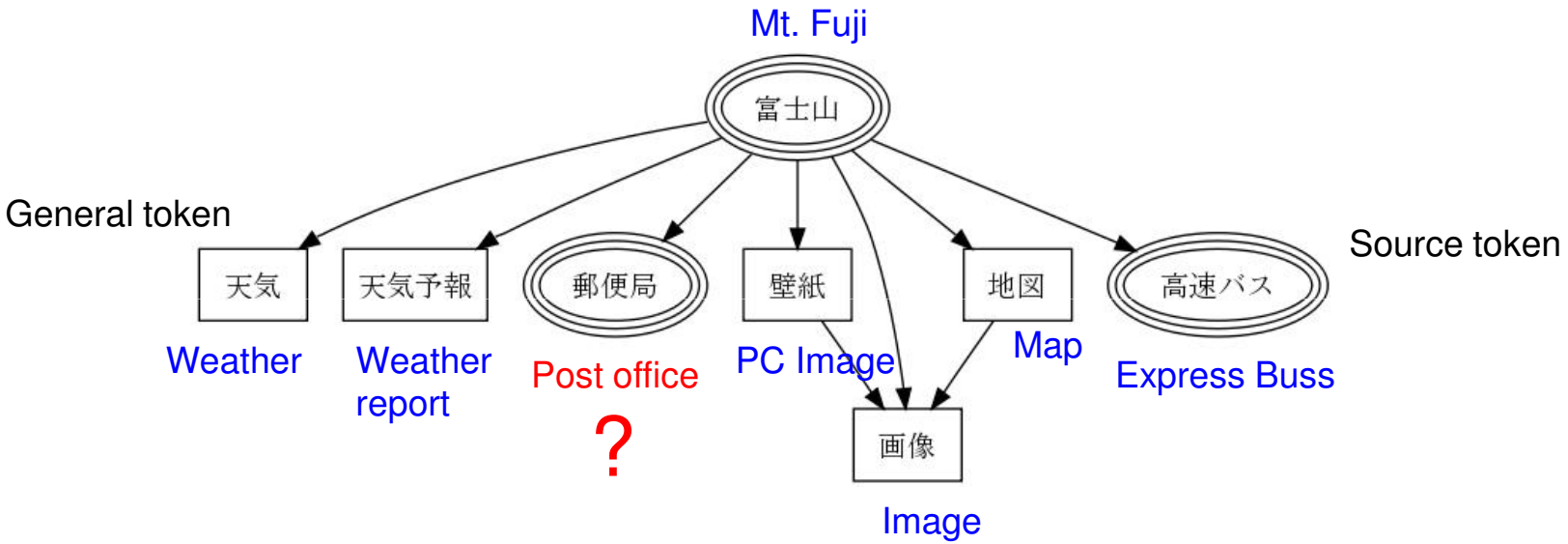
$$N_T = 400, N_R = 50$$

- 66 similar token pairs are found, including
  - Various synonym expressions of Beijing Olympic
    {五輪, 北京五輪}, {五輪, 北京オリンピック}
  - Similar free movie sites in Japan
    {ニコニコ動画, youtube}
  - Similar places for summer vacation
    {沖縄, Hawaii} (Okinawa, Japan and Hawaii)
  - Other pairs have contextual similarity

- Findings
  - Distributional Similarity provides two different types of words
    - Similarity of words themselves (synonym)
    - Similarity of contexts in use

Mt. Fuji

- Topic graph shows Mt. Fuji as a sightseeing place in summer

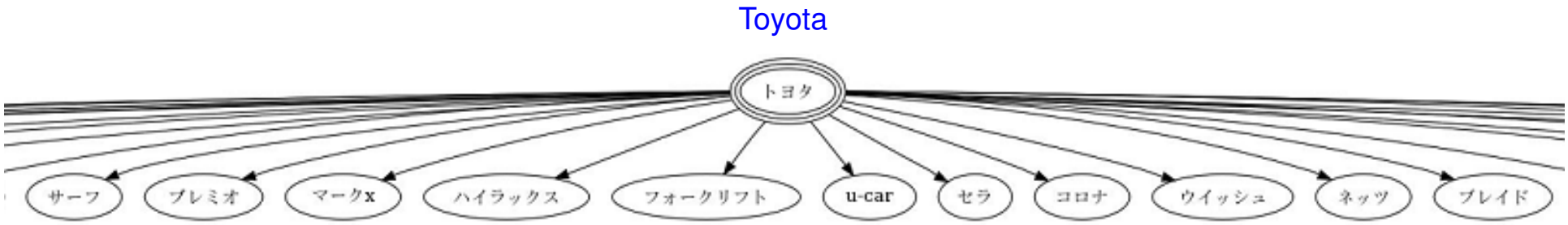


- Because in summer many climbers send out mail at the post office in the top of Mt.Fuji.

YAHOO! JAPAN

Toyota

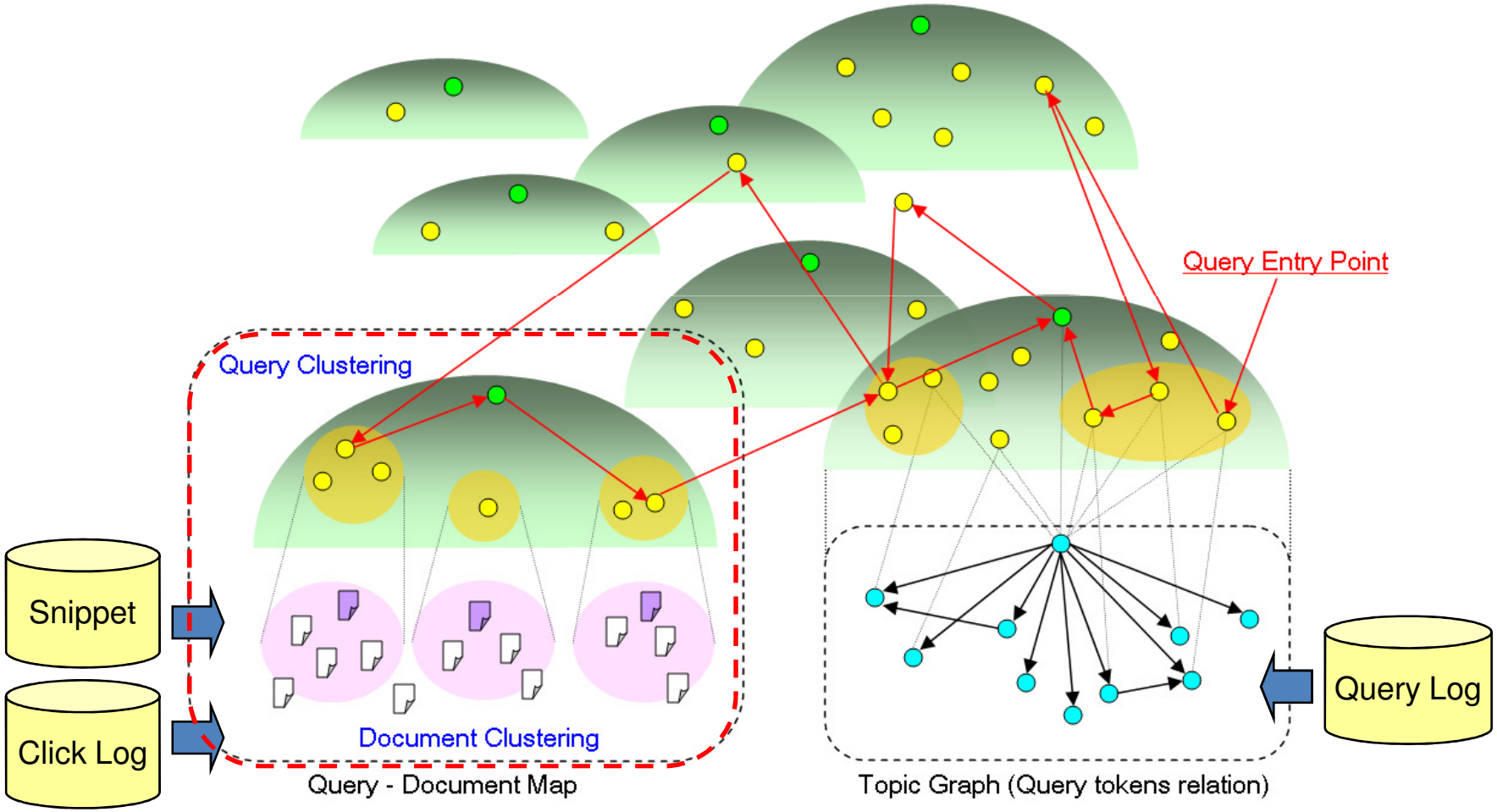
- Topic graph shows Toyota's many car lineup.
- "中古車" (used car) has links to some of specific four cars;
  Prius, Vitz, Hi-Ace, Aristo



YAHOO! JAPAN

Building Query-Document Map

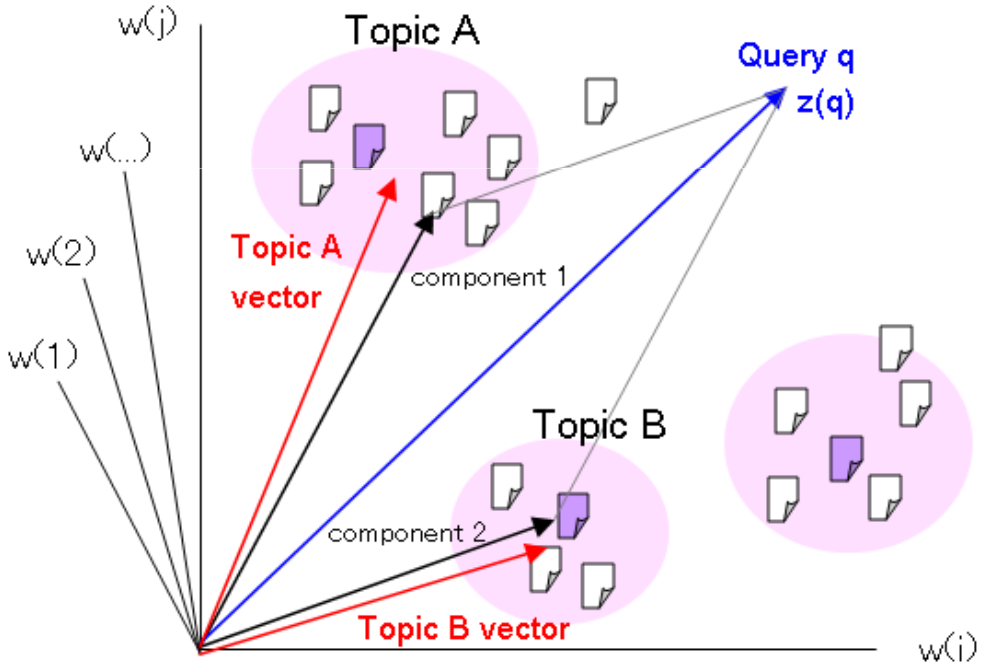
- Based on topic graph, click log and snippet in the results, we can build query-document map



YAHOO! JAPAN

Document Clustering and Query Sense Decomposition

- Clicked documents of a topic T_i can be clustered by using word vectors extracted from snippet.
  - The evaluation was done very well.
- Query sense decomposition can be done by using click distribution on URLs in the search results. (Seems it would work well.)



Query Vector decomposition on Query-Document Map

(Multi-dimensional word space based on semantic distance)

YAHOO! JAPAN

Chapter 3

Recommendation Strategy

Model-based approach for adaptation to
user behaviors and dynamic lexical sense change

Towards Information Supply in 4th generation search

- Information supply means (My thoughts)
  - Easiest information access
  - Help people discover variable and unreachable information
- So, search engines need to know
  - Meanings of query
  - User's (query) intent
  - Topics and trends, etc.

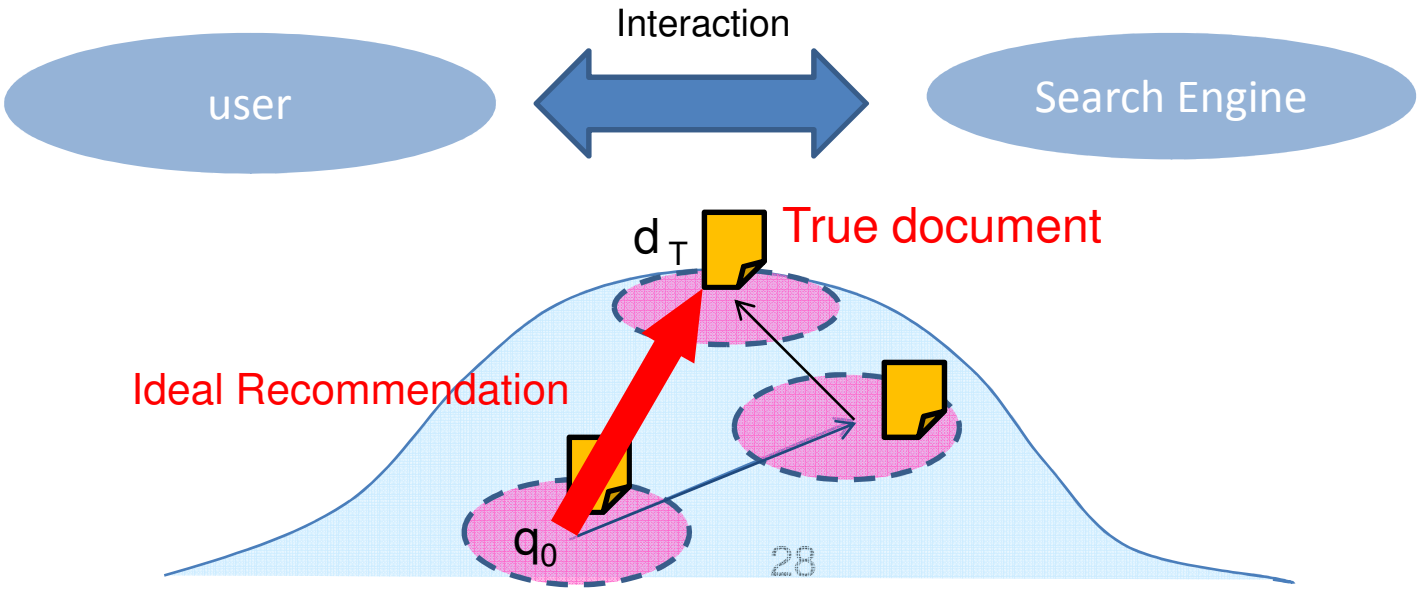
  } Analysis and Discovery
- Then, search engines also need
  - Good recommendation strategy

How to help user?

Relevance feedback (Idea of IR field) doesn't work always

1. User's query is not always correct
   Search behavior is interaction (Query refinement process)
2. User's click is not always correct
   Some clicks are just in examinations

From observation of user behavior,
we had better focus on global model, not personal preference

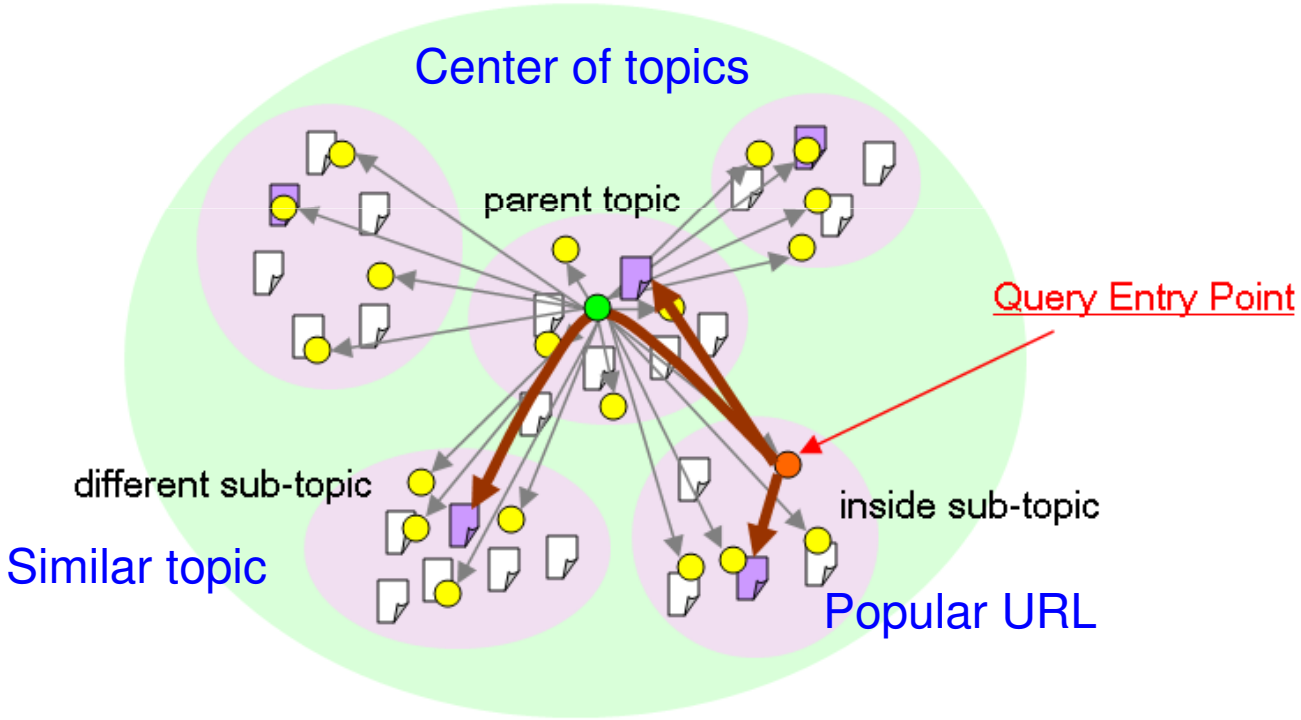


YAHOO! JAPAN

How to Recommend What?

The idea is “Control and Navigation” using the map for recommendation

- Three possible recommendation strategies on the document-query map
- Serendipity recommendation could be realized in this framework

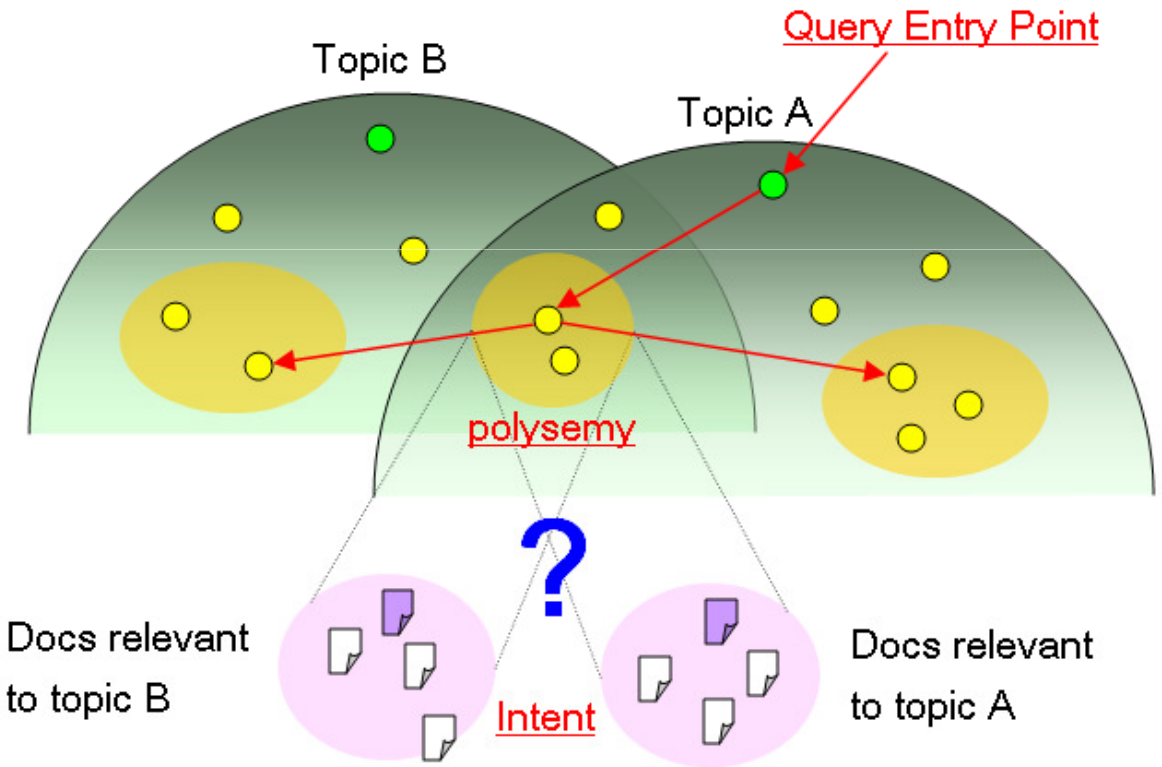


Examples of Different Recommendation Strategies

YAHOO! JAPAN

Applying to Query Sense Disambiguation

- If a query has different meanings, using the map we can do
  - Topic Identification (disambiguation)
  - Query intent analysis on document-query map (***Model***)



Query Sense Disambiguation (An approach to Lexical Polysemy)

YAHOO! JAPAN

Chapter 4

Conclusion and Future Work

Towards Information Supply Search Engine

- Presented one of ideas to realize 4th generation search
- Discussed how to capture topics and trends emerged from collective user behaviors
- Proposed some algorithms for Analysis & Discovery, and typical recommendation (information supply) strategy
- Next step
  - Integration with Machine Learning and Knowledge Acquisition framework
  - Improvement of the theory

Thank you!

YAHOO! JAPAN