Web Document Clustering

using

Semantic Link Analysis

SOMJIT ARCH-INT, Ph.D.

Semantic Information Technology Innovation (SITI) LAB
Department of Computer Science, Faculty of Science,
Khon Kaen University, THAILAND

IAWTIC'2005
28-30 November 2005, Vienna, AUSTRIA
Outline

- **Introduction**
- **Proposed Model**
  1. Semantic (Concept) Clustering & Scoring
  2. Semantic Document Clustering & Scoring
  3. Searching & Ranking
- **Experiment & Results**
- **Conclusion**
Introduction

- Traditional Web Search Approach
  - Most search engines use keyword-based algorithms (i.e., algorithms based on TF/IDF schema)
    - Keyword-based matching
    - Often return irrelevant documents
    - Always ignore the data meaning in a document.
  - Thus, the proper criteria for a search engine in evaluating document retrieval must be refined and addressed.
Two Related Tasks (1/2)

- A richness context of information:
  - Context which expresses the meaning of Web documents must be specified and annotated
  - A set of semantic-based relevant documents will be provided.

→ However, searching on the Web requires the quality of contents called authoritativness or trusted sources of correct information.
Two Related Tasks (2/2)

Link analysis ranking:

- A number of hyperlink analysis methods are proposed to derive a quality measure for the information on the Web.
- The assumption of the link-based analysis depends on the number of links of a Web page in order to identify a page with high or low quality.
  - Intuitively, a link from page $p$ to page $q$ denotes an endorsement for the quality and authoritativeness of page $q$.

- A simple counting of the number of links to a Web page should not be taken into account, due to the fact that not all the citations have the same authority.

- Moreover, the use of link-based analysis to rank Web documents leads to the problem of “circular contribution effect” due to the hubness (e.g., a message board page).
Proposed Model

- Hybrid Model:
  - Using both context **information richness** and **link analysis ranking** methods to categorize relevant and authoritative documents into a single cluster.

- Two basic steps for semantic document clustering:
  - Semantic Information Extraction
    - Extracts the context of information to be the concept for Web documents using WordNet, resulting in a set of concepts.
  - Semantic Link Analysis.
    - Relevant documents are clustered into a single cluster.
    - Clustered documents is ranked using link analysis method based on the semantic clustered information.

**.... Semantic Information Extraction**
1. Semantic Information extraction

- Keywords extracting and clustering
- Semantic cluster scoring
1. Semantic Information extraction

- Extract the contexts of Web document features into concepts (keyword representatives).

- Groups related keywords relied on a common concept as a single semantic cluster (sc).

- The ranking method is evaluated based on the common concept of a strong interconnectivity.
Definition 1: Semantic Clustering (SC)

For any extracted keyword (ki) in the document d, keywords will be categorized into a single semantic cluster if and only if these keywords have a common concept.

- Match the meaning of incoming keywords with the synonym set (synset_id) in WordNet and
- cluster the synonymous keywords into a single semantic cluster ($sc_i$).
Definition 1: Semantic Clustering (Cont.)

Let

- $K_d = \{k_1, k_2, \ldots, k_m\}$ be the set of extracted keywords of a Web document $d$,
- $WN = \{id_1, id_2, \ldots, id_T\}$ be the set of synonym set identifiers in the WordNet, and
- $SC = \{sc_1, sc_2, \ldots, sc_N\}$, semantic clusters, represent the concept space in the domain.
  - Each $sc_i$ composes of a set of keywords.
Definition 2: Semantic Cluster Weighting (SC-W)

A semantic cluster weight, $SC-W_{(sc,d)}$, of a cluster $sc$ belonging to a document $d$ is the average weight computed using keyword weights of the cluster $sc$.

\[
SC-W_{sc(j),d} = \frac{\sum_{i \in sc_j} w_{i,d}}{|sc_j|}
\]

where

- $|sc_j|$ = the number of keywords within $d$ that have the same semantic cluster $j^{th}$.
- $w_{i,d}$ = weight of the keyword $i^{th}$ in the document $d$.
- $SC-W_{sc(j),d}$ is the average weight of cluster $j^{th}$.

The $SC-W_{sc(j),d}$ is a rank of the document $d$ based on the semantic cluster, $sc$. 
2. Semantic Link Analysis

- **Semantic Document Clustering**
  - The purpose of semantic document clustering is to logically categorize Web documents which contain a common semantic cluster into a single semantic document cluster.
2. Semantic Link Analysis (Cont.)

- **Definition 3: Semantic-based Document Clustering (S-DC)**

  Any semantic cluster \( (sc_i) \) as defined in Definition 1 in the domain will be categorized into a single semantic document cluster if these semantic clusters have a common concept and have at least one link (in- or out-link).

Let

- \( G = (V, E) \) be a Web directed graph consisting of a set of Web documents (pages) and a set of links among pages.

This research considers only two types of links, that is, *out-link* and *in-link*, by omitting *loop link*. 
2. Semantic Link Analysis (Cont.)

- **Semantic Document Scoring**
  - Semantic document scoring is determined using the notion of the page ranking and the document similarity.
  - Given a document $u$ which contains a semantic cluster representative (concept) $sc$, the significant level of the document $u$ is computed by employing the link analysis as shown in equation (6).

$$S^{(i+1)}(u,sc) = k \sum_{v \in B_u} \frac{S^{(i)}(v,sc)}{N_v} \times SIM_{sc}(v,u)$$  \hspace{1cm} (6)

where

- $S^{(i)}(v,sc)$ is the significant score computed in the $i^{th}$ loop and
- $SIM_{sc}(v,u)$ is the semantic similarity (score) between document $v$ and $u$ based on the common concept $sc$. 

3. Searching

- **Searching Process**

  Given the index scores of semantic-based concept \( (sc-w_{sc,d}) \) and Web documents score \( (s_{(sc,d)}) \) in a database.

  Let
  - \( Q = \{q1, q2, ..., qm\} \) be the set of query consisting of \( m \) sub-queries (keywords) and
  - \( W_{i,Q} \) be the weight of the keyword \( i^{th} \) in \( Q \) that are computed using (TF/IDF scheme).

- **Searching evaluates the semantic similarity between \( Q \) and document representatives in the database.**
  - The semantic similarity, \( SIM(Q,d) \), between a particular document, \( d \), and the query, \( Q \), will be measured in terms of cosine coefficient as equation (8).

\[
SIM(Q,d) = \frac{\sum_{\forall i \in Q} W_{sc,i} \cdot S_{sc,d}}{\sqrt{\sum_{\forall i \in Q} W_{sc,i}^2 \cdot \sqrt{\sum_{\forall sc \in d} S_{sc,d}^2}}}
\]

where
- \( S_{sc,d} \) is the semantic score of the document \( d \) based on semantic cluster \( sc \).
Experiment

Experiment Sample: using kku.ac.th domain
- Collected 24,599 Web documents
- These Web documents were generated into 27,864 semantic document clusters from 2,091,641 keywords

Precision Measure

Precision = \frac{\text{No. of relevant documents retrieved}}{\text{No. of relevant documents retrieved} + \text{No. of irrelevant documents retrieved}} \quad (9)

Experiment Method

- The experiment randomly selected 375 queries and divided them into 5 sets, 75 samples for each.
- The retrieved documents determined using equation (8) were classified into 5 categories, namely, Top-1, Top-5, Top-10, Top-20, and Top-30
  - Top-1 retrieved documents means that the precision considers only the first retrieved document to be a relevant document, whereas other documents are counted to be both relevant or irrelevant documents.


- **Experiment Results**

  - Compare the quality of searching between the proposed *semantic-based document clustering* method and the *traditional search* method (TF/IDF).
  - The measurement of precision is computed as shown in the graph.

    ![Graph showing precision comparison between TF/IDF and Semantic methods.]

    - The proposed method was able to outperform the TF/IDF method up to 9.31% on average.
    - The performance of the proposed method is rather stable whereas the TF/IDF scheme performs in reverse tendency manner between degrees of retrieved documents and the relevant documents.
Conclusion

- **The principle contribution:**
  - the semantic document clustering and
  - semantic-based link analysis methods.

- The approach uses WordNet as the assistant mechanisms of extracting and clustering.

- The experimental results reaffirm that the proposed method outperforms the traditional search method (keyword-based method) on average.
Acknowledgement

- This work is supported by
  - The National Research Council of THAILAND (NRCT).
  - Department of Computer Science, Faculty of Science, KKU.

http://www.cs.kku.ac.th
Discussion