WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques

Ricardo Campos 1, 2, Gaël Dias 2, Célia Nunes 2

1 Instituto Politécnico de Tomar
   Tomar, Portugal

2 Centre of Human Language Technology and Bioinformatics
   Universidade da Beira Interior
   Covilhã, Portugal

The 2006 IEEE / WIC / ACM International Conference on Web Intelligence, Hong Kong - China, December 20, 2006

Agenda

→ Introduction
→ Purposes
→ Related Work
→ Proposed Solution
→ Results
→ Future Work
→ Conclusions
**Introduction**

**Purposes**

**Related Work**

**Proposed Solution**

**Results**

**Future Work**

**Conclusions**

---

**Difficult task in selecting the best relevant documents**

**No semantic analysis**

**Ambiguity problem and the synonymous problem**

---

**Contextualization**

To tackle these drawbacks

---

web meta search engine called **WISE**

---

- web content mining techniques
- statistical methodologies
- we group documents on the fly

---

we increase the quality of the presented results, hierarchically grouped and concept disambiguated.
The representation of the documents

In order to represent the contents of the retrieved documents two approaches are mainly considered:

1. **Vector Space Model**
   - The set of relevant terms is calculated under TF-IDF measure.

2. **Shared n-grams**
   - The set of relevant terms are the terms shared by more than one document.

The different approaches, also distinguish themselves by the match of the two following items:

- **Simple words or phrases**;
- **flat or hierarchical clustering**.

<table>
<thead>
<tr>
<th>Simple Words</th>
<th>Hierarchical Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Flat Clustering</strong></td>
<td><strong>Hierarchical Clustering</strong></td>
</tr>
</tbody>
</table>
The architecture of WISE is simple

WISE is a web search interface system

The system returns a page with a set of clusters and their associated key concepts

Bellow each keyword there exist a list of urls, in one or more of these groups

The Architecture of WISE

Our algorithm is composed of five steps:

1. Search results gathering;
2. Selection of relevant web pages;
3. Document parsing for Phrase Extraction;
4. Document parsing for Key Concept Extraction;
5. Hierarchical Clustering and Labeling.
The first step is formalized in Equation 1:

\[ R = r_s(d_i \mid q), \quad \forall i, 1 \leq i \leq n \]  

(1)

where \( q \) is the query, \( d_i \) a document, \( r_s \) a function of a search engine \( s \), that calculates the relevance between \( q \) and \( d_i \).

So in the second step, we apply a selection function over \( R \), defined in Equation 2:

\[ R' = \text{select}(R) \]  

(2)

where \text{select} is the algorithm that selects the most relevant documents over the set of web pages search results;
**Introduction**

*Selection of Relevant Web Pages*

where a relevant document is:

- any absolute URL retrieved as such (i.e. the web domain);
- any URL which number of occurrences\(^1\) exceeds the average relevance defined in next equation.

\[
\text{average_relevance} = \frac{\text{# returned_urls}}{\text{# different_absolute_urls}} = \frac{4}{3} = 1.33
\]

\(^1\) calculated under the sum of all URLs that share the same absolute URL.

---

Ricardo Campos, Gaël Dias, Célia Nunes  
WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques

---

**Conclusion**

We catch the \(n\) best web pages of the absolute URLs

```
www.3dfon.com
www.3dfon.com/wes/99.html
www.3dfon.com/soa.html
```

Our system turns out to be more robust
**Document Parsing for Phrase Extraction**

*Hong Kong, Web Intelligence, Luis Figo, left wing or to score a goal*

In the third step, each document of $R'$ is parsed to extract phrases. This step is defined in Equation 4:

$$S = \text{senta}(d_j), d_j \in R', j = 1..., \#R'$$

(4)

where $S$ is the set of documents in which all phrases replace the relevant sequences of single words;

**Document Parsing for Key Concept Extraction**

*each phrase of a document is represented by a set of 8 properties:*

- the importance of the phrase in the document;
- the relationship between the phrase and the query in the document collection.

*We use WebSpy, which is defined as a set of decision trees (C5.0 class) previously trained with oversampling to avoid data sparseness;*
The representation of the documents is given by a vector of key concepts containing its most relevant scored phrases, as formalized in Equation 5, we represent each document \( d_j \) of \( S \), as a vector of key concepts \( kc_j = (kc_{j1}, kc_{j2}, \ldots, kc_{jm}), j=1,\ldots, #R' \)

\[
kc_j = \text{webspy}(d_j, q), d_j \in S
\]  

where \( m \) is the number of key concepts for \( d_j \) retrieved by Webspy taking into consideration the query \( q \).

Each element of the key concept vector, also known as key concept \( (kc) \), is a flat cluster (only one-level partitioning of the data) with a list of related urls.

At this step we achieve flat clustering.

\[
C_0 = \{d_j | d_j \in S \} \quad \text{and} \quad d_{rt}(d_j, kc_j) \]
**Introduction**

**Purposes**

**Related Work**

**Proposed Solution**

<table>
<thead>
<tr>
<th>Search results gathering</th>
<th>Results</th>
<th>Future Work</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection of relevant web pages</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hierarchical Clustering and Labelling</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Search results gathering**

**Selection of relevant web pages**

**Hierarchical Clustering and Labelling**

**Document Parsing for Key Concept Extraction**

Flat clustering still presents a major drawback:

- the user feels lost with so many information;
- the set of results share different concepts, but disposed in undistinct way (genuinely ambiguity problem unsolved).

To simplify the user’s search process, we propose a soft hierarchical structure.

**Hierarchical Clustering and Labelling**

The organization of the results into clusters eases the user’s navigation.

To achieve it, we first need to represent each document $d_v$, $v=1,...,#C_h$, of each cluster $C_h$, as a vector of key concepts $k_{C_v} = (k_{C_v1}, k_{C_v2},..., k_{C_vn})$, as formalized in equation 7:

$$k_{C_v} = \text{webspy}(d_v | k_{C_h}), d_v \in C_h$$  \hspace{1cm} (7)

where $t$ is the number of key concepts for $d_v$ retrieved by Webspy taking into consideration the query $k_{C_h}$.
Then, we apply the PoBOC algorithm to obtain the set of hierarchical clusters HC:

\[ HC = \text{poboc}(\{\text{sim}(k_{c_v}, k_{c_{v'}}) | v \neq v' \land v, v' = 1, \ldots, \#C_s\}) \]  

(8)

where sim is the cosine similarity measure defined in 9:

\[ \text{cos}(k_{c_v}, k_{c_{v'}}) = \frac{k_{c_v} \cdot k_{c_{v'}}}{|k_{c_v}| \cdot |k_{c_{v'}}|} \]  

(9)

By applying the PoBOC algorithm to the flat clusters, we propose a disambiguation methodology, as key concepts with different meanings should be gathered in different clusters;

By doing so, the possible different meanings of the query would be evidenced;

PoBOC is a soft clustering algorithm based on graph theory, that can be used on the fly, as it does not depend on any input parameter, and has shown encouraging results compared to other clustering algorithm when applied to text data;
Hierarchical Clustering and Labelling

The process of clustering does not end with the creation of coherent clusters;

The label of the hierarchical clusters is obtained by $HC'$, applying the label function

\[ HC' = label(HC) \]  

using a simple heuristics that chooses:

1. the key concept that occurs more often in the vectors representing each flat clustering in the cluster;

2. taking into account the sum of scores in case of ties.

Conclusions

The system is able to deal with word mistake

(1) The labels are semantically descriptive

(2) The system is able to deal with word mistake

Centre of Human Language Technology and Bioinformatics - University of Beira Interior - Portugal - [h l t g . d i . u b i . p t ]

Ricardo Campos, Gaël Dias, Célia Nunes
## Bigger coverage results; Overlap

(3) The coverage of our results is much bigger

(4) Overlap

Centre of Human Language Technology and Bioinformatics - University of Beira Interior - Portugal - [huliti@ubi.pt]

| Ricardo Campos, Gaël Dias, Célia Nunes | WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques |

## Language-Independence

(5) Language-Independent

Centre of Human Language Technology and Bioinformatics - University of Beira Interior - Portugal - [huliti@ubi.pt]

| Ricardo Campos, Gaël Dias, Célia Nunes | WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques |
Concept Disambiguation

(7) Ability to deal with concept disambiguation:

Cluster José-António-Camacho → Benfica Club
Cluster PS → Benfica Neighborhood
Cluster Universitários → Transports/Housing in Benfica

Evaluate the system;

Experimental results

- Correctness of the clusters
- Quality and descriptiveness of the labels
- Concept disambiguation and language-independence

but formal evaluation is needed
**Improve Performance; WebWarehouse**

It is clear that organizing web search results fastens the user browsing process.

The architecture and the proposed algorithms are the solution to one of the biggest problems with search engines actually deal with:

- The return of quality results
  - flexible and real-world adaptable structure (language/domain/topic independent),
  - automatic
  - organized
  - concept disambiguate.

Integrate in the future a web warehouse to save the key concepts of the documents.

Integrate an on-line demo (wise.di.ubi.pt).

Centre of Human Language Technology and Bioinformatics - University of Beira Interior - Portugal - [hultig.di.ubi.pt]
WISE has a strong theoretical basis, but its architecture is simple;

In particular it is based on:

- **an algorithm** ignoring less relevant documents and increasing relevant ones;
- statistical parameter-free **phrase extraction** to define concepts enhancing document understanding;
- **web content mining techniques** to semantically represent document contents;
- a **soft clustering** parameter-free algorithm to organize results into a hierarchy of concepts and a classical labeling process.

Thanks for your attention!