Document Clustering for Mediated Information Access - The WebCluster Project -

David J. Harper, Mourad Mechkour and Gheorghe Muresan
School of Computer and Mathematical Sciences
The Robert Gordon University, Aberdeen

Sponsor: Ubilab, Union Bank of Switzerland, Zurich

Contents

• Motivation - problems with searching the WWW

• Main idea - mediated access through a clustered collection

• Document clustering in WebCluster

• Users and scenarios of use

• Evaluation

• Ongoing and future work

• Conclusions
**WebCluster - Motivation**

- **Information Need**
  - (within some subject domain, e.g. NDT)

  - Query
  - Infoseek

  **Gulfs**
  - information need <-> query
  - WWW (search engine) <-> WWW (subject domain)
  - unstructured WWW <-> structured subject domain (non-WWWW)

---

**WebCluster - Mediated Access**

- **Source Collection (domain)**
  - Information Need

  - Structured Source Collection

  - Query
  - Infoseek
**Clustering a Document Collection : an Example**

Goals for clustering

- Reveal the structure of the domain and its main concepts
- Group similar documents

---

**Mediating Access to the WEB : a Concrete Example**

1. **Clustering**
2. Cluster-based search + browsing
3. query
4. Result
5. **Private Document Collection**
The Mediated Access Assumptions

- During a search session, the user is interested in a relatively narrow domain
- The domain has a semantic / topical structure
- A specialised collection covering the domain of interest is available
- The structure of the collection is either explicit (the collection is categorized) or implicit / inherent (it can be ‘discovered’ by clustering)
- The collection is homogeneous (and so is the indexing)

Mediated Access to the WWW

The search - a two stage process:

Stage 1 : Browse / cluster-based search the Source Collection
Stage 2 : Access to the relevant document from the Target Collection
WebCluster scenario #1

Name
Explicit mediated access to the WWW

Targeted users
Experienced searchers

Principle
• Cluster based search of the source collection
• Display and browse the structure of the source collection
• Search the WWW using query based on the cluster representative

WebCluster scenario #2

Name
Implicit mediated access to the WWW

Targeted users
Naive / beginner searchers

Principle
• Cluster based search of the clustered source document collection
• Display and browse the documents in a chosen cluster:
  • Sample documents: from the source collection and from the WWW pre-populating the clusters
  • Whole set of documents produced by automatically querying the WWW using a query based on the cluster representative
Informal experiment

Objectives

- Test the users’ reaction to the mediated access concept
- Test the user satisfaction regarding the functionality of the system, and the relevance of the documents retrieved
- Formative usability testing (the volunteers were not only experienced searchers, but also had experience in evaluating IR systems)
- Comparison of user generated queries vs. system generated queries

Experimental procedure

- Subjects received introduction to the system
- Task assigned: “You are a trainee in a newspaper. You support the journalists by providing information for their articles.”
- Sample topics:
  - The history of the Brazilian debt crisis
  - How are the quotas for growing coffee set and controlled on a world-wide basis?
- Source collection: a Reuters collection of newspaper articles
- Steps followed by users:
  - Formulate a query and record it
  - Browse source collection, select ‘best’ cluster, edit query generated by system, submit it to the search engine
  - Submit to the search engine the initial, self-generated query
  - Compare results of the two searches
- Methodology: direct observation and think-aloud protocol
Informal experiment

Results

- Users found the mediation useful for unfamiliar topics

- The two-step process was questioned when:
  - query formulation was considered easy, for a familiar topic
  - the documents of the source collection were considered sufficient to cover the information need

- The system proposed good query terms, that increased recall

- Users not always good at recognising ‘good’ query terms

- The system proposed bad query terms, that decreased precision

- More functionality required

On-going research

- The user-interface
  - Better metaphor for the user interface - ClusterBook
    - Book - dynamic table of contents, index
    - Encyclopaedia
  - Structured feedback
  - Colour model
  - Better support for query formulation
    - Multiple-topic searching

- Evaluation
  - User-generated query vs. system-generated query
  - Ranked search vs. mediated search

- Better algorithms for cluster representatives and query generation
Future work

• Integration with a Web browser

• Meta-searching for the appropriate source collection

• Brokers/proxies for distributed source collections

• Multi-user system

• Integrate different media and different indexing methods

The WebCluster project - Conclusions

Objective

• Access very large, heterogeneous document collections

Main idea

• Use for mediation a small, homogeneous, domain-specific document collection, clustered in order to reveal the domain semantic / topical structure

Services

• Help the users understand the main concepts of the domain, how they are represented in the document collection, and view some very relevant documents

• Help the user in the query formulation process

• Visualization tools for exploring the search results
Architecture
WebCluster - Applications

CD-ROM based collections
  • structured access to the collection itself
  • mediated access to WWW (via CD-ROM)

Mediated access via hierarchically structured information sources. Examples are: via large structured report (e.g. government reports), via structured collection of information (e.g. encyclopaedia), via intranet

Multimedia information access
  • cluster multimedia source, e.g. annotated photographs
  • mediated access to other photographs (not annotated)