Motivation

- **Focused crawling** identifies Web pages relevant to a selected topic
- **Automatic information extraction** automatically extracts structured data from unstructured or semi-structured Web pages
- **Set expansion** uses Web pages content to automatically completes a given small set of examples

Automatic Information Extraction

- We extract structured data from web pages automatically by first inferring the underlying structure (wrapper)
- We use unsupervised systems for generating wrapper: RoadRunner [3], ExAlg [2]
  - Both of these systems only exploits the structure of web pages without prior knowledge of the content
  - ObjectRunner [1] incorporates structured object definition as a way for the user to describe the targeted data

SEAL

- SEAL [5] generates wrapper for each page and extract candidates from the same page
  - A wrapper is defined as a pair of maximally long common left context and maximally long right context
  - It uses an heterogeneous graph model, node as entities (seeds, document, wrapper, mention) and edge as the relations between nodes
  - It performs lazy walk on the graph to measure the similarity between two nodes(similar to PageRank)
  - The rank of any entities in the graph can be calculated using this weight

Example

- Result from SEAL using the examples “Yves Rocher” and “L’Oreal” as seeds for a search of cosmetic brand names: “Maybelline”, “Biotherm”, “Guerlain”, “Clinique”, “Garnier”, “Lancome”, “Boscia”, “Dior”, etc.

Research Objectives

- Integrate and extend approaches for focused crawling, automatic information extraction and set expansion
- Design and implement a practical system able to extract structured data from the web
- Design a ranking mechanism based on the analysis of the heterogeneous graph of objects (websites, web pages, wrapper, data elements, and concepts in ontologies)

The system will also able to rank websites, web pages and ontology elements

Set Expansion Problem

- Extract elements of a particular semantic class from a given data source
  - For example: given the seeds (Barrack Obama, George Bush, Bill Clinton) extract more element of the particular semantic class (US Presidents) from the web
- Three steps framework:
  - Fetch relevant documents: find occurrences of the seeds in the collection of documents
  - Construct patterns and extract candidates: using regular expression to construct patterns and from these patterns extract the candidates
  - Rank candidates: use some kind of ranking mechanism on candidates, usually variations of PageRank [4]

Bibliography