This presentation is an introduction to the problem of search results clustering. SRC can be considered part of Web Mining, a dynamically growing branch of Information Retrieval. In this paper we give a definition of the problem, and give a brief positioning of the field comparing it to the classic document clustering. Results from an experimental clustering system are presented and problems are highlighted. Finally we propose some future research directions, listing open issues with existing algorithms.
The roadmap

- What is clustering about?
- What is Search Result Clustering about?
  - Problem definition
  - Differences between SRC and traditional clustering
- (Quick) review of existing methods and algorithms
- Clustering of search results in Polish: Carrot system
- Future research directions
- Questions
What is clustering about?

A problem of clustering is perceived differently, depending on the context of application. This humorous example will explain what is understood under the term of clustering in Search Results Clustering.

I will use UML to explain this concept. This slide presents a set of UML Actors (let’s think of them as Objects). Can you distinguish any clusters (groups) of them in this slide?
What is clustering about?

The two possible clusters are small and large actors.

Being a software engineer, I know UML is a highly extensive language. I decided to extend the notation with…
What is clustering about?

...facial expression of actors. Now, what clusters are present in this group of objects?
What is clustering about?

...there are many possible options, of course, of one them is to split the set of actors into happy and unhappy ones (those unhappy actors are company clients, those happy ones – developers).

To add an element of maths to the UML notation, I present yet another UML extension: Actor’s thinking mode.
What is clustering about?

We have three different thinking modes: rigid logic, no logic and fuzzy logic (please try to decide which one is which). What clusters can be distinguished now?
What is clustering about?

One possible split is to divide actors according to the thinking mode they’re using…
Slide 9

What is clustering about?

... but one could as well combine features and say one cluster is composed of logically thinking actors, and the other of happy actors. The decision is solely a matter of choice of attributes (features) of objects in the set being clustered.
What is clustering about?

- Discovering similarities in a set of objects
- Decrease the amount of information and express it in a concise way
- Clusters structure is not known in advance (classification), however similarity/measure is usually given a priori
- The obvious: structure of clusters highly depends on object features/ criteria considered in an algorithm

Thus, we have come to the conclusions presented above.

The most important features of clustering in Search Results Clustering are _decreasing_ the volume of information presented to the user, and choosing the most intuitive features organizing the set of search results into groups.
The roadmap

- What is clustering about?
- What is Search Result Clustering about?
  - Problem definition
  - Differences between SRC and traditional clustering
- (Quick) review of existing methods and algorithms
- Clustering of search results in Polish: Carrot system
- Future research directions
- Questions
What is Search Results Clustering about?

- Internet is a massive source of unstructured information
- Keyword-based search engines return thousands of links, but lack explanation of these results
- Advanced IR methods and NLP are still not efficient enough
- An alternative presentation of results from search engines is required

In addition to this slide, let’s say that an alternative presentation scheme is needed for ranked-list-based search engines. However, a totally opposite approach is also required – current web search engines are document-retrieval engines, while in the future we would prefer to have information-retrieval systems as well (those capable of answering questions based on the knowledge extracted from the Web, not only returning relevant Web sites).
Search Results Clustering - definition

Search Results Clustering is about efficient identification of meaningful, thematic groups of documents in a search result and their concise presentation.
A query „king” yielded the documents presented in above. Can you identify what the results are about? What topics do they describe?
This is easier, right? We immediately see meaningful topics like „Stephen King”, „Martin Luther King”, „King County College”. Clustering of results proved to be an efficient way of explaining the results.
Search Results Clustering - process

- **INPUT**
  - N links to documents (a search result)
    - (0<N<~400), each composed of an URL, an optional title and a *snippet*
- **ASSUMPTIONS**
  - There exists a logical structure of topics in the result set
- **OUTPUT**
  - A set of clusters representing topics, possibly organized in a hierarchical, overlapping structure
- **ALGORITHM?**

*A simplified requirements of a Search Results Clustering algorithm.*
There are many classic clustering algorithms. This slide explains why most of them are not applicable to the problem of SRC.
The roadmap

- What is clustering about?
- What is Search Result Clustering about?
  - Problem definition
  - Differences between SRC and traditional clustering
- (Quick) review of existing algorithms
- Clustering of search results in Polish: Carrot system
- Future research directions
- Questions
Review of existing algorithms - STC

- **Suffix Tree Clustering**, Oren Zamir, Oren Etzioni, used in Grouper system
  - *Document feature:* shared, descriptive phrases
  - *Algorithm speed:* online, incremental, Ukkonen’s suffix tree construction
  - *Full phrases form names of clusters*
  - *Performs poorly in languages with less strict word order, or lack of explicit word separators (Chinese)*

An in-depth explanation of STC can be found in Oren Zamir’s doctoral thesis.
Review of existing algorithms - SHOC

- SHOC (Semantic, hierarchical, online clustering), Dell Zhang, Yisheng Dong
  - Document feature: shared, descriptive phrases
  - Algorithm speed: logarithmic (!), suffix array used instead of a tree
  - A more complex clusters merging technique

- Results promising (?) WICE system.

Very little is known about SHOC. The paper describing it is to be published as of 01/07/2002.
### Proprietary algorithms

- **Vivisimo, commercial clustering engine**
  - Algorithm: unknown, a heuristic utilizing phrases
  - Generates very intuitive clusters for English and quite good ones for Polish (without terms stemming!)
  - A very good comparison reference
- **Infonetware, commercial**
  - Flat structure of clusters
- **Excavio, commercial**
  - Data Mining – based?

*Vivisimo yields the best clustering results among all services known to the author. However, they don’t reveal the details of the algorithm.*
### The roadmap

- What is clustering about?
- What is Search Result Clustering about?
  - Problem definition
  - Differences between SRC and traditional clustering
- (Quick) review of existing algorithms
- Clustering of search results in Polish: Carrot system
- Future research directions
- Questions
Carrot – Clustering of Polish Search results

- **Motivation**
  - Verify STC’s applicability to Polish
  - A testbed for new algorithms
- **Polish language customizations:**
  - quasi-stemming
  - stop-words

Carrot was a project meant to implement and verify the applicability of STC to the Polish language. More details can be found on author’s website: http://www.cs.put.poznan.pl/dweiss
Carrot – project conclusions

- As expected, STC did quite poorly on Polish sentences
- Hard to define proper threshold values
- A need for quality stemming is evident
- Pruning junk words besides stop words is needed

- Still, the system yield some promising results. Several ideas pending for Carrot II (codename: cluster-master, pol. klajster-majster)
The roadmap

- What is clustering about?
- What is Search Result Clustering about?
  - Problem definition
  - Differences between SRC and traditional clustering
- (Quick) review of existing algorithms
- Clustering of search results in Polish: Carrot system
- Future research directions
- Questions
Future research directions

- Discovering new potential similarity measures, search result attributes useful for the clustering process
- Heuristics instead of exact algorithms? Rule-based cluster creation?
- Employing NLP-based methods?
- Employing Fuzzy Sets?
- What will Google do?!
Thank you

Questions?

Further questions, please e-mail to:
dawid.weiss@cs.put.poznan.pl