Web Information Retrieval

Dipl.-Inf. Christoph Carl Kling
Exercises WebIR

ask questions!

WebIR@c-kling.de
Random jump with $e = 0.3$

( normal jump with $(1-e) = 0.7$ )

→ the markov chain becomes ergodic!
Exercise:

Calculate

\[ a_1 = a_0 P \]
\[ a_2 = a_1 P \]
\[ \ldots \]
\[ a_6 = a_5 P \]

Will a converge? And why?
Hubs and Authorities (or HITS) algorithm

Authority scores  \( x_1 = (0.25, 0.25, 0.25, 0.25) \)

Hub scores  \( y_1 = (0.25, 0.25, 0.25, 0.25) \)
Exercise:

Calculate

\[ x_2, y_2, \ldots x_4, y_4 \]

and normalise \( x \) and \( y \) after each step so that each vector sums to 1 (that is, the values can be interpreted as probabilities).
Exercise:

“Hubs and Authorities” - unlike Page Rank - does not take the whole web graph, but only a subset. Describe how the Hubs and Authority algorithm works!
Exercise:

What do we do with the Page Rank scores?
Page Rank

What do we do with the Page Rank scores?

1) We calculate the PR for every website in our corpus

2) A user asks the query q

3) We calculate a relevance-score for documents based on q

4) We (e.g.) multiply the relevance score with the PR score for the final ranking
Exercise:

How do we manipulate the system?

100% original V14GR4 - Buy now!!!
Exercise:

How do we manipulate the system?

1) Manipulate the text based relevance score
2) Manipulate the Page Rank
In clinical VIAGRA practice, VIAGRA doctors personally assess patients in order to diagnose, treat, and prevent disease VIAGRA using VIAGRA clinical judgment. The doctor-patient relationship typically begins an VIAGRA interaction with an examination of the patient's VIAGRA medical history and medical VIAGRA record, followed a medical interview[4] VIAGRA and a physical examination. Basic VIAGRA diagnostic medical devices (e.g. stethoscope, tongue depressor) are typically used. After examination VIAGRA for signs and interviewing for VIAGRA symptoms, the doctor may VIAGRA order medical tests (e.g. blood tests), take a VIAGRA biopsy, or prescribe pharmaceutical VIAGRA drugs or other therapies. Differential diagnosis methods help to rule out conditions VIAGRA based on the information provided. During the encounter, properly informing the VIAGRA patient of all relevant facts is an important VIAGRA part of the relationship VIAGRA and the development of trust. The medical VIAGRA encounter is then documented VIAGRA in the medical record, which is a legal document in many jurisdictions.[5] Followups may VIAGRA be shorter but follow the same general VIAGRA procedure.
2) Manipulate Page Rank score

4: originalv14gr4.com
2) Manipulate Page Rank score

4: originalv14gr4.com
2) Manipulate Page Rank score

4: originalv14gr4.com
Classification

Given data items with labels, predict the label for a new item.

Why do we need classification methods in WebIR?

Which classification methods do we know?
Classification

Given data items with labels, predict the label for a new item.

Which classification methods do we know?

- Naive Bayes ("Probabilistic Retrieval")
- K-nearest neighbours
Exercise:

What is clustering?
Exercise:

What is clustering?

- Grouping of Data so that:
  → Data within a group are similar
  → Data of different groups are dissimilar

→ Reduction in complexity of data
Exercise:

What is clustering good for in the context of WebIR?
Exercise:

What is clustering good for in the context of WebIR?

- Faceted browsing
- Diversification of search results
- Preselecting relevant documents
K-Means

What's the Input / Output?
K-Means

Input
- Data points in a space where we have a distance measure (e.g. n-dimensional vector space with Euclidean distance) and where we can calculate means
- Number of clusters (K)
K-Means

**Input**
- Data points in a space where we have a distance measure (e.g. n-dimensional vector space with Euklidean distance)
- Number of clusters (K)

**Output**
- Cluster assignments of data points
- Cluster means (called “centroids”)
K-Means

Algorithm

1) Randomly assign data points to the K clusters

2) Calculate cluster centroids

3) Assign data points to the cluster with the closest centroid

4) Repeat 2) and 3) till centroids don't change or change only a little
Clustering & Classification

K-Means

Problems

- Normalisation of input data
- Outliers
- Clustering not unique
- Clustering dependant of initialisation → “bad” clustering
- Empty clusters
- 1-element clusters
- non “circular” data
- different data densities
Thank you!
Questions?
Classification Evaluation

- Accuracy
- Precision
- Recall
- F-Measure