UNIVERSITÄT DES SAARLANDES

Personalisation

Seminar on

Unlocking the Secrets of the Past: Text Mining for Historical Documents

Sven Steudter

Domain

- Museums offering vast amount of information
- **But:** Visitors receptivity and time limited
- **Challenge:** seclecting (subjectively) interesting exhibits
- Idea of mobile, electronice handheld, like PDA assisting visitor by :
  1. Delivering content based on observations of visit
  2. Recommend exhibits
- Non-intrusive ,adaptive
  user modelling technologies used

Prediction stimuli

UNIVERSITÄT DES SAARLANDES

- Different **stimuli**:
  - Physical **proximity** of exhibits
  - Conceptual simularity (based on **textual** describtion of exhibit)
  - Relative sequence other visitors visited exhibits (**popularity**)
- Evaluate relative impact of the different factors => seperate stimuli
- Language based models simulate visitors thought process

Experimental Setup

UNIVERSITÄT DES SAARLANDES

- Melbourne Museum, Australia
- Largest museum in Southern Hemisphere
- Restricted to Australia Gallery collection, presents history of city of Melbourne:
  - Phar Lap
  - CSIRAC
- **Variation of exhibits** : can not classified in a single category

Experimental Setup

UNIVERSITÄT DES SAARLANDES

- Wide range of modality:
  - Information plaque
  - Audio-visual enhancement
  - Multiple displays interacting with visitor
- Here: NOT differentiate between exhibit types or modalities
- Australia Gallery Collection exists of 53 exhibits
- Topology of floor: open plan design => **no predetermined sequence** by architecture

Resources

UNIVERSITÄT DES SAARLANDES

- Floorplan of exhibition located in 2. floor
- Physical distance of the exhibits
- Melbourne Museum web-site provides corresponding web-page for every exhibit
- Dataset of 60 visitor paths through the gallery, used for:
  1. Training (machine learning)
  2. Evaluation

Predictions based on Proximity and Popularity

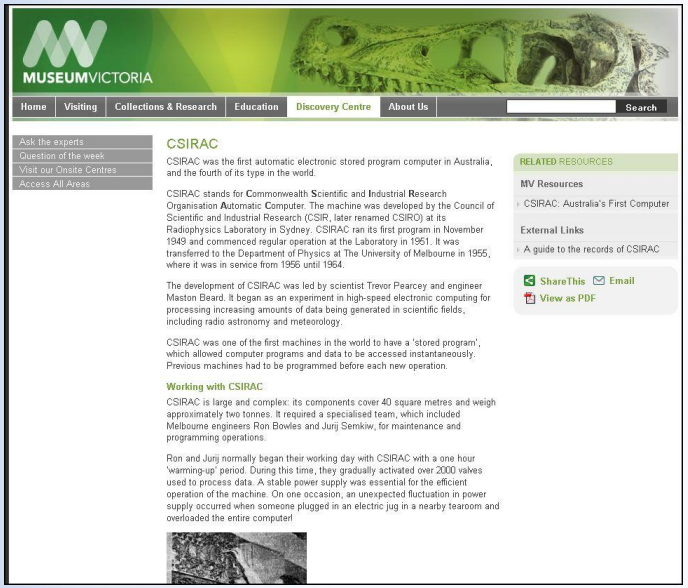
UNIVERSITÄT DES SAARLANDES

- Proximity-based predictions:
  - Exhibits ranked in order of physical distance
  - Prediction: **closest not-yet-visited exhibit** to visitors current location
  - In evaluation: baseline
- Popularity-based predictions:
  - Visitor paths provided by Melbourne Museum
  - **Convert paths** into matrix of transitional probabilities
  - Zero probabilities removed with Laplacian smoothing
  - Markov Model

Text-based Prediction

UNIVERSITÄT DES SAARLANDES

- Exhibits related to each other by information content
- Every exhibits web-page consits of:
  1. Body of text describing exhibit
  2. Set of attribute keywords
- Prediction of most similar exhibit:
  - Keywords as queries
  - Web-pages as document space
- Simple term frequency-inverse document frequency, tf-idf
- Score of each query over each document normalised




```
<head><title>
        Museum Victoria: CSIRAC
</title><meta name="Title" content="Museum Victoria: CSIRAC" />
<meta name="Description" content="The oldest electronic computer in the world, CSIRAC, is a part of Museum Victoria's History and Technology Collection." />
<meta name="Keywords" content="Museum,Melbourne,Victoria,computer,CSIRAC,Maston Beard,Trevor Pearcey,electronic,data,CSIR,CSIRO,Australia,Australian,technology,invention,Ron Bowles,Jurij Semkiw,program,
<meta name="Creator" content="Museum Victoria" />
<meta name="Created" content="Wed, 01 Jan 2003 00:00:00 GMT" />
<meta name="Last-Modified" content="Thu, 01 May 2008 15:21:01 GMT" />
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
```

WSD

UNIVERSITÄT DES SAARLANDES

- Why make visitors connections between exhibits ?
- Multiple simularities between exhibits possible
- Use of Word Sense Disambiguation:
  - Path of visitor as sentence of exhibits
  - Each exhibit in sentence has associated meaning
  - Detemine meaning of next exhibit
- For each word in keyword set of each exhibit:
  - WordNet similarity is calculated against each other word in other exhibits

WordNet Similarity

UNIVERSITÄT DES SAARLANDES

- Similarity methods used:
  - Lin (measures difference of information content of two terms as function of probability of occurence in a corpus)
  - Leacock-Chodorow (edge-counting: function of length of path linking the terms and position of the terms in the taxonomy)
  - Banerjee-Pedersen (Lesk algorithm)
- Similarity as sum of WordNet similarities between each keyword

$$\frac{\sum_{k_1 \in K_1} \sum_{k_2 \in K_2} WNsim(k_1, k_2)}{|K_1||K_2|}$$

- Visitors history may be important for prediction
- Latest visited exhibits higher impact on visitor than first visited exhibits

Evaluation: Method

UNIVERSITÄT DES SAARLANDES

- For each method two tests:
  1. Predict next exhibit in visitors path
  2. Restrict predictions, only if pred. over threshold
- Evaluation data, aforementiened 60 visitor paths
- 60-fold cross-validation used, for Popularity:
  - 59 visitor paths as training data
  - 1 remaining path for evaluation used
  - Repeat this for all 60 paths
  - Combine the results in single estimation (e.g average)

Evaluation

UNIVERSITÄT DES SAARLANDES

Accuracy: Percentage of times, occured event was predicted with highest Probability

BOE: Bag of Exhibits: Percentage of exhibits visited by visitor, not necessary in order of recommendation

BOE is, in this case, identical to precision

Method	BOE	Accuracy
Proximity (baseline)	0.270	0.192
Popularity	0.406	0.313
Tf·Idf	0.130	0.018
Lin	0.129	0.039
Leacock-Chodorow	0.116	0.024
Banerjee-Pedersen	0.181	0.072
Popularity - Tf·Idf	0.196	0.093
Popularity - Lin	0.225	0.114
Popularity - Leacock-Chodorow	0.242	0.130
Popularity - Banerjee-Pedersen	0.163	0.064
Proximity - Tf·Idf	0.205	0.084
Proximity - Lin	0.180	0.114
Proximity - Leacock-Chodorow	0.220	0.151
Proximity - Banerjee-Pedersen	0.205	0.105
Proximity - Popularity	0.232	0.129

Single exhibit history

Evaluation

UNIVERSITÄT DES SAARLANDES

Method	BOE	Accuracy
Proximity (baseline)	0.270	0.192
Popularity	0.406	0.313
Tf·Idf	0.130	0.018
Lin	0.129	0.039
Leacock-Chodorow	0.116	0.024
Banerjee-Pedersen	0.181	0.072
Popularity - Tf·Idf	0.196	0.093
Popularity - Lin	0.225	0.114
Popularity - Leacock-Chodorow	0.242	0.130
Popularity - Banerjee-Pedersen	0.163	0.064
Proximity - Tf·Idf	0.205	0.084
Proximity - Lin	0.180	0.114
Proximity - Leacock-Chodorow	0.220	0.151
Proximity - Banerjee-Pedersen	0.205	0.105
Proximity - Popularity	0.232	0.129

Method	Threshold	Precision	Recall	F-score
Proximity	0.03	0.271	0.270	0.270
Popularity	0.06	0.521	0.090	0.153
Tf·Idf	0.06	0.133	0.122	0.128
Lin	0.01	0.129	0.129	0.129
Leacock-Chodorow	0.01	0.117	0.117	0.117
Banerjee-Pedersen	0.01	0.182	0.180	0.181
Popularity - Tf·Idf	0.001	0.176	0.154	0.164
Popularity - Lin	0.0005	0.383	0.316	0.348
Popularity - Leacock-Chodorow	0.0005	0.430	0.349	0.385
Popularity - Banerjee-Pedersen	0.001	0.236	0.151	0.184
Proximity - Tf·Idf	0.001	0.189	0.174	0.181
Proximity - Lin	0.0005	0.239	0.237	0.238
Proximity - Leacock-Chodorow	0.0005	0.252	0.250	0.251
Proximity - Banerjee-Pedersen	0.0005	0.182	0.180	0.181
Proximity - Popularity	0.001	0.262	0.144	0.186

Single exhibit history

without threshold

with threshold

Evaluation

UNIVERSITÄT DES SAARLANDES

Method	BOE	Accuracy
Proximity (baseline)	0.270	0.192
Popularity	0.406	0.313
Tf·Idf	0.130	0.018
Lin	0.129	0.039
Leacock-Chodorow	0.116	0.024
Banerjee-Pedersen	0.181	0.072
Popularity - Tf·Idf	0.196	0.093
Popularity - Lin	0.225	0.114
Popularity - Leacock-Chodorow	0.242	0.130
Popularity - Banerjee-Pedersen	0.163	0.064
Proximity - Tf·Idf	0.205	0.084
Proximity - Lin	0.180	0.114
Proximity - Leacock-Chodorow	0.220	0.151
Proximity - Banerjee-Pedersen	0.205	0.105
Proximity - Popularity	0.232	0.129

Single exhibit history

Method	BOE	Accuracy
Proximity	0.066	0.0
Popularity	0.016	0.0
Tf·Idf	0.033	0.0
Lin	0.064	0.0
Leacock-Chodorow	0.036	0.0
Banerjee-Pedersen	0.036	0.0

Visitors history enhanced

Conclusion

UNIVERSITÄT DES SAARLANDES

- Best performing method: Popularity-based prediction
- History enhanced models low performer, possible reason:
  - Visitors had no preconceived task in mind
  - Moving from one impressive exhibit to next
- History **here** not relevant, current location more important
- Keep in mind:
  - Small data set
  - Melbourne Gallery (history of the city) perhabs no good choice

BACKUP

tf-idf

UNIVERSITÄT DES SAARLANDES

- Term frequency – inverse document frequency
- Term count=number of times a given term appears in document
- Number n of term t_i in doument d_j
- In larger documents term occurs more likely, therefor normalise

$$\mathrm{tf}_{\mathrm{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

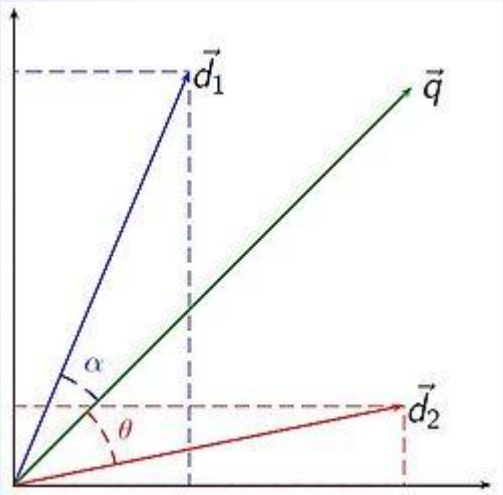
- Inverse document frequency, idf, measures general importance of term
- Total number of documents,
- Divided by nr of docs containing term

$$\mathrm{idf}_{\mathrm{i}} = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

tf-idf: Similarity

- Vector space model used
- Documents and queries represented as vectors
- Each dimension corresponds to a term
- Tf-idf used for weighting
- Compare angle between query an doc

$$sim(d_j, q) = \frac{\mathbf{d_j} \cdot \mathbf{q}}{\|\mathbf{d_j}\| \, \|\mathbf{q}\|} = \frac{\sum_{i=1}^{t} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} * \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$



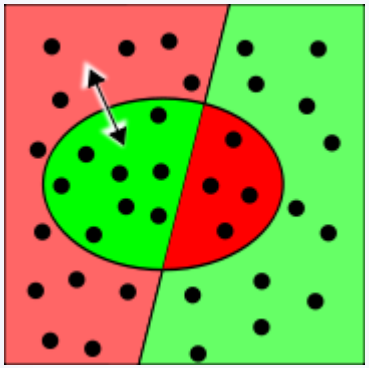
WordNet similarities

UNIVERSITÄT DES SAARLANDES

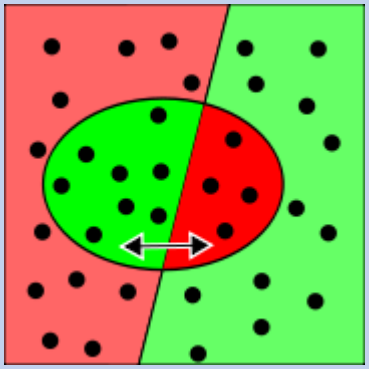
- Lin:
  - method to compute the semantic relatedness of word senses using the information content of the concepts in WordNet and the 'Similarity Theorem'
- Leacock-Chodorow:
  - counts up the number of edges between the senses in the 'is-a' hierarchy of WordNet
  - value is then scaled by the maximum depth of the WordNet 'is-a' hierarchy
- Banerjee-Pedersen, Lesk:
  1. choosing pairs of ambiguous words within a neighbourhood
  2. checks their definitions in a dictionary
  3. choose the senses as to maximise the number of common terms in the definitions of the chosen words.

Precision, Recall

Recall: Percentage of relevant documents with respect to the relative number of documents retrieved.



$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$



Precision: Percentage of relevant documents retrieved with respect to total number of relevant documents in dataspace.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

F-Score

- F-Score combines Precision and Recall
- Harmonic mean of precision and recall

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$