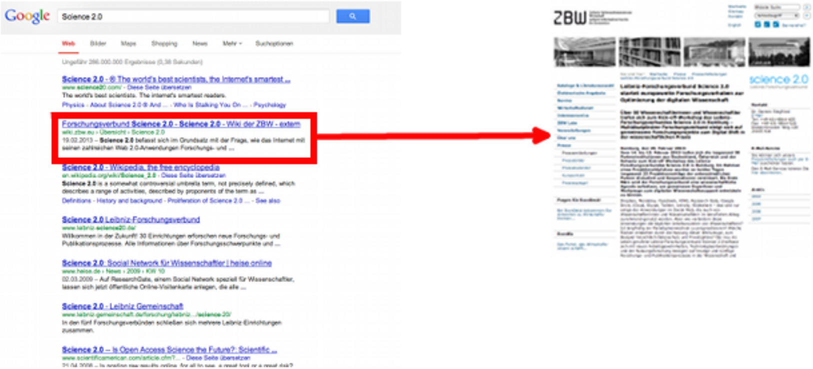
A Contribution to Rating and Recommendation Systems: Concepts, Development and Evaluation

Oliver Diestel

March 4, 2013

Current Situation (Just an assumption)

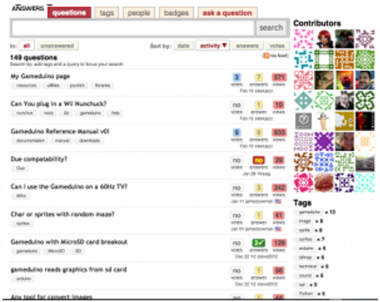


User visits the zbw.eu website and searchs for the information that he would like to have. Leaves the website.

Possible Situation

User visits the website, sees an interesting question from another user, reads the question and the answer, hopefully asks more questions. Goal: Increase the user interaction on the website.

Q/A System



General Idea: Use a Question/Answer System to make the information retrieval process public. Use these public questions and answers to increase the user interaction on the website.

Beginning

What do we have at the beginning?
Website, Users and Questions

Rating

- click
- active time on page
- individual click actions

Rating

- click
- active time on page
- individual click actions

What do we have now?
Questions, users and the information, which user likes which question

Tagging

Find words that describe the question.

Tagging

Find words that describe the question.

- STW Thesaurus for Economics

Tagging

Find words that describe the question.

- STW Thesaurus for Economics
- Stemming: Compute the stem of a word.

Tagging

Find words that describe the question.

- STW Thesaurus for Economics
- Stemming: Compute the stem of a word.
  Example rules
  "ies" $\rightarrow$ "y" und "s" $\rightarrow$ ""
  Examples
  "libraries" $\rightarrow$ "library" and "Wikis" $\rightarrow$ "Wiki"

Tagging

Find words that describe the question.

- STW Thesaurus for Economics
- Stemming: Compute the stem of a word.
  Example rules
  "ies" $\rightarrow$ "y" und "s" $\rightarrow$ ""
  Examples
  "libraries" $\rightarrow$ "library" and "Wikis" $\rightarrow$ "Wiki"
- Computing Levenshtein distance: Calculate the distance between two words.
  Example:
  Libraries, Library
  Distance: 3

Tagging

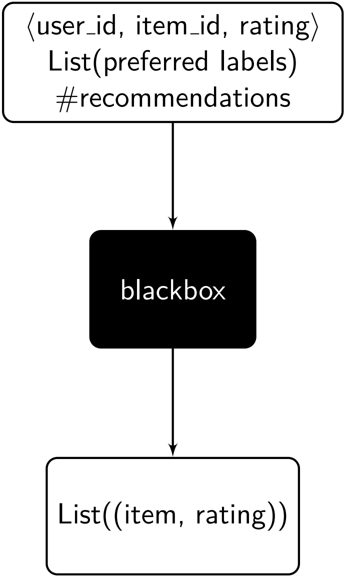
Find words that describe the question.

- STW Thesaurus for Economics
- Stemming: Compute the stem of a word.
  Example rules
  "ies" $\rightarrow$ "y" und "s" $\rightarrow$ ""
  Examples
  "libraries" $\rightarrow$ "library" and "Wikis" $\rightarrow$ "Wiki"
- Computing Levenshtein distance: Calculate the distance between two words.
  Example:
  Libraries, Library
  Distance: 3

What do we have now?
Information about which user likes which question and one or more preferred label(s) that describe(s) the subject of the question

Input/Output

- Item: id, name, one or more preferred label(s)
- User: id, name



Item-Based Algorithm

	Item1	Item2	Item3	Item4	Item5
User1	5	3	4	4	?
User2	3	1	2	3	3
User3	4	3	4	3	5
User4	3	3	1	5	4
User5	1	5	5	2	1

Item-Based Algorithm

	Item1	Item2	Item3	Item4	Item5
User1	5	3	4	4	?
User2	3	1	2	3	3
User3	4	3	4	3	5
User4	3	3	1	5	4
User5	1	5	5	2	1

Cosinus Similarity

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|}$$

Example

$$\text{sim}(\text{Item5},\text{Item1}) = \frac{3 \cdot 3 + 5 \cdot 4 + 4 \cdot 3 + 1 \cdot 1}{\sqrt{3^2+5^2+4^2+1^2} \cdot \sqrt{3^2+4^2+3^2+1^2}} = 0.99$$

Predictions

Prediction

User u, Item p, Rating $r_{u,p}$

$$pred(u,p) = \frac{\sum_{i \in ratedItems(u)} sim(i,p) \cdot r_{u,i}}{\sum_{i \in ratedItems(u)} sim(i,p)}$$

Predictions

Prediction

User u, Item p, Rating $r_{u,p}$

$$pred(u,p) = \frac{\sum_{i \in ratedItems(u)} sim(i,p) \cdot r_{u,i}}{\sum_{i \in ratedItems(u)} sim(i,p)}$$

Example

sim(Item5, Item1) $= 0.99$
sim(Item5, Item2) $= 0.74$
sim(Item5, Item3) $= 0.72$
sim(Item5, Item4) $= 0.94$

$$pred(User1, I5) = \frac{0.99 \cdot 5 + 0.74 \cdot 3 + 0.72 \cdot 4 + 0.94 \cdot 4}{0.99 + 0.74 + 0.72 + 0.94} = 4.07$$

What do we have now?

Matrix with less empty fields than before.

	Item1	Item2	Item3	Item4	Item5
User1	5	3	4	4	4
User2	3	1	2	3	3
User3	4	3	4	3	5
User4	3	3	1	5	4
User5	1	5	5	2	1

Singular Value Decomposition

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

Create a SVD with the matrix $M = U \cdot \Sigma \cdot V^t$

Singular Value Decomposition

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

Create a SVD with the matrix $M = U \cdot \Sigma \cdot V^t$

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix}$$

corresponds to the row
vectors of matrix m

Singular Value Decomposition

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

Create a SVD with the matrix $M = U \cdot \Sigma \cdot V^t$

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix}$$

corresponds to the row vectors of matrix m

$$\begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix}$$

diagonal matrix with $\sigma_{ii} > 0$ and $\sigma_{ii} \geq \sigma_{i+1i+1}$

Singular Value Decomposition

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix}$$

Create a SVD with the matrix $M = U \cdot \Sigma \cdot V^t$

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix}$$

corresponds to the row vectors of matrix m

$$\begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix}$$

diagonal matrix with $\sigma_{ii} > 0$ and $\sigma_{ii} \geq \sigma_{i+1i+1}$

$$\begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{pmatrix}$$

corresponds to the column vectors of matrix m

Low Rank Approximation of M

$$\begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{pmatrix}$$

- Derive from Σ the matrix Σ_k (with k new rank of M) formed by replacing σ_{ii} with $i > k$ by zeros.
- Compute and output $M_k = U \cdot \Sigma_k \cdot V^T$ as the rank-k approximation to M.

Low Rank Approximation of M

$$M_2 = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{pmatrix}$$

- Derive from Σ the matrix Σ_k (with k new rank of M) formed by replacing σ_{ii} with $i > k$ by zeros.
- Compute and output $M_k = U \cdot \Sigma_k \cdot V^T$ as the rank-k approximation to M.

Low Rank Approximation of M

$$M_2 = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{pmatrix}$$

$$M_2 = \begin{pmatrix} u_{11} \cdot \sigma_{11} & u_{12} \cdot \sigma_{22} & u_{13} \cdot 0 \\ u_{21} \cdot \sigma_{11} & u_{22} \cdot \sigma_{22} & u_{23} \cdot 0 \\ u_{31} \cdot \sigma_{11} & u_{32} \cdot \sigma_{22} & u_{33} \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{pmatrix}$$

Low Rank Approximation of M

$$M_2 = \begin{pmatrix} u_{11} & u_{12} & \cancel{u_{13}} \\ u_{21} & u_{22} & \cancel{u_{23}} \\ u_{31} & u_{32} & \cancel{u_{33}} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{pmatrix}$$

$$M_2 = \begin{pmatrix} u_{11} \cdot \sigma_{11} & u_{12} \cdot \sigma_{22} & u_{13} \cdot 0 \\ u_{21} \cdot \sigma_{11} & u_{22} \cdot \sigma_{22} & u_{23} \cdot 0 \\ u_{31} \cdot \sigma_{11} & u_{32} \cdot \sigma_{22} & u_{33} \cdot 0 \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \\ v_{13} & v_{23} & v_{33} \end{pmatrix}$$

Example SVD

$$\begin{array}{cccccc} & \text{Item1} & \text{Item2} & \text{Item3} & \text{Item4} & \text{Item5} \\ \text{User1} & 5 & 3 & 4 & 4 & 4 \\ \text{User2} & 3 & 1 & 2 & 3 & 3 \\ \text{User3} & 4 & 3 & 4 & 3 & 5 \\ \text{User4} & 3 & 3 & 1 & 5 & 4 \\ \text{User5} & 1 & 5 & 5 & 2 & 1 \end{array} =$$

Example SVD

$$
\begin{array}{lccccc}
 & \text{Item1} & \text{Item2} & \text{Item3} & \text{Item4} & \text{Item5} \\
\text{User1} & 5 & 3 & 4 & 4 & 4 \\
\text{User2} & 3 & 1 & 2 & 3 & 3 \\
\text{User3} & 4 & 3 & 4 & 3 & 5 \\
\text{User4} & 3 & 3 & 1 & 5 & 4 \\
\text{User5} & 1 & 5 & 5 & 2 & 1
\end{array}
=
$$

$$
\begin{pmatrix}
0.544178 & 0.0875457 & 0.303701 & -0.598262 & -0.496039 \\
0.330582 & 0.270665 & 0.155794 & -0.281254 & 0.845033 \\
0.517601 & 0.04377 & 0.429477 & 0.737011 & -0.0503875 \\
0.438285 & 0.373564 & -0.800903 & 0.129864 & -0.100232 \\
0.366854 & -0.881822 & -0.239996 & -0.0541261 & 0.165165
\end{pmatrix}
\cdot
$$

$$
\begin{pmatrix}
16.499 & 0 & 0 & 0 & 0 \\
0 & 4.93905 & 0 & 0 & 0 \\
0 & 0 & 2.58239 & 0 & 0 \\
0 & 0 & 0 & 1.20841 & 0 \\
0 & 0 & 0 & 0 & 0.511218
\end{pmatrix}
\cdot
$$

$$
\begin{pmatrix}
0.452438 & 0.336841 & 0.410894 & -0.456436 & -0.55197 \\
0.403968 & -0.531238 & -0.483027 & 0.210155 & -0.526419 \\
0.43523 & -0.601119 & 0.481498 & -0.122706 & 0.449817 \\
0.463447 & 0.282982 & -0.586236 & -0.401113 & 0.447854 \\
0.477391 & 0.403612 & 0.149459 & 0.756011 & 0.12371
\end{pmatrix}^{T}
$$

Diagram of SVD

	Item1	Item2	Item3	Item4	Item5
User1	5	3	4	4	4
User2	3	1	2	3	3
User3	4	3	4	3	5
User4	3	3	1	5	4
User5	1	5	5	2	1

$$\begin{pmatrix} 0.544178 & 0.0875457 \\ 0.330582 & 0.270665 \\ 0.517601 & 0.04377 \\ 0.438285 & 0.373564 \\ 0.366854 & -0.881822 \end{pmatrix}$$

Diagram of SVD

	Item1	Item2	Item3	Item4	Item5
User1	5	3	4	4	4
User2	3	1	2	3	3
User3	4	3	4	3	5
User4	3	3	1	5	4
User5	1	5	5	2	1

$$\begin{pmatrix} 0.544178 & 0.0875457 \\ 0.330582 & 0.270665 \\ 0.517601 & 0.04377 \\ 0.438285 & 0.373564 \\ 0.366854 & -0.881822 \end{pmatrix}$$

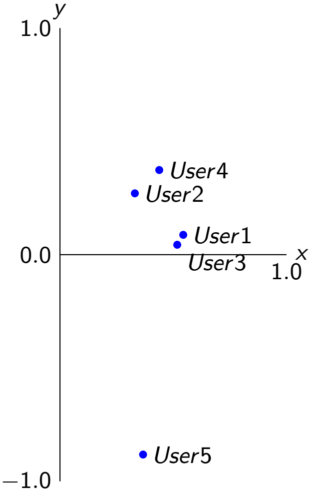
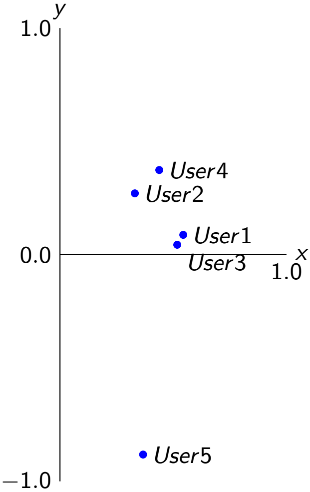


Diagram of SVD

	Item1	Item2	Item3	Item4	Item5
User1	5	3	4	4	4
User2	3	1	2	3	3
User3	4	3	4	3	5
User4	3	3	1	5	4
User5	1	5	5	2	1

$$\begin{pmatrix} 0.544178 & 0.0875457 \\ 0.330582 & 0.270665 \\ 0.517601 & 0.04377 \\ 0.438285 & 0.373564 \\ 0.366854 & -0.881822 \end{pmatrix}$$



Recommendations for User

- Take the k most similar users SU
- Take the top k unknown items (with the right context) from each user $\in$ SU
- Use cosinus similarity for rating weighting

Is this a good approach?

- We could have calculated the SVD directly out of the original matrix. Would we get a similar result?

Is this a good approach?

- We could have calculated the SVD directly out of the original matrix. Would we get a similar result?
- Is it better if we use the V matrix from the SVD to calculate the item similarity and update the SVD afterwards?

Is this a good approach?

- We could have calculated the SVD directly out of the original matrix. Would we get a similar result?
- Is it better if we use the V matrix from the SVD to calculate the item similarity and update the SVD afterwards?
- Is it better if we just recommend similar items from the items the user already likes?

Is this a good approach?

- We could have calculated the SVD directly out of the original matrix. Would we get a similar result?
- Is it better if we use the V matrix from the SVD to calculate the item similarity and update the SVD afterwards?
- Is it better if we just recommend similar items from the items the user already likes?
- Should we use cosinus similarity between the users in the first place.

Evaluation

The test data: I use real data from the zbw econdesk. However, we do not have real user ratings for this data.

- Every data gets a random quality value from 1 (rather negative rating) to 5 (rather positive rating).
- Furthermore, I generate 1000 test users. These users will have a rating preferation. So a user might be a person that rates an item more positive or more negative.
- I will try to evaluate the algorithms with this test data. I might use a movie db as well for this.

Software Architecture

Scala: Finagle twitter framework

Every part of the software is a service.

The tagger, the rating algorithm and the recommendation algorithms can be used as an individual software.

Timetable

Start: 18.12.2012
End: 18.06.2013
1st month(18.01): Theory
2nd month(18.02): Theory + Technology
3rd month(18.03): Implementation
4th month(18.04): Implementation + First writings
5th month(18.05): Final thesis
Personal Goal
Technology implemented: 31.03.2013
Diplom Thesis ready: 01.05.2013