

# Evaluation Strategies and Methods

Christian Körner

[christian.koerner@tugraz.at](mailto:christian.koerner@tugraz.at)

Graz University of Technology, Austria

# Agenda for Today

- Scenario
- Important Notes
- Case Studies
  - Types of Measures
    - Qualitative
    - Quantitative
- Limitations
- Take Home Messages(s)

# Scenario

Using the knowledge acquired in this course you have developed a new method for knowledge acquisition.

## 2 Questions arise:

- How do you show that your effort is better than existing work?
- If no such work exists (“Pioneer” status): How do you know that your work simply “works”?

## Important Notes / 1

Without evaluation there is no proof that your discovery/work is correct and significant

A good evaluation design takes time to be constructed

Evaluation helps you to support your claims / hypotheses

## Important Notes / 2

It is often not possible to evaluate everything! - Only excerpts!

Creativity is needed

Evaluation techniques are not carved in stone

This is not a complete list of evaluation techniques (by far)

# Types of Methods

Qualitative Measures

Quantitative Measures

# Qualitative Evaluation

## With the help of empirical studies

- Human subjects who evaluate samples of the results
- The more people you have the merrier
- Agreement between test subjects is important

# Quantitative Evaluation

With the help of proven measures

comparing the your effort to a golden standard

“compete” with a lower bound - baseline

## 2 case studies

### Evaluation of the Goal Prediction Interface:

- Example for qualitative evaluation

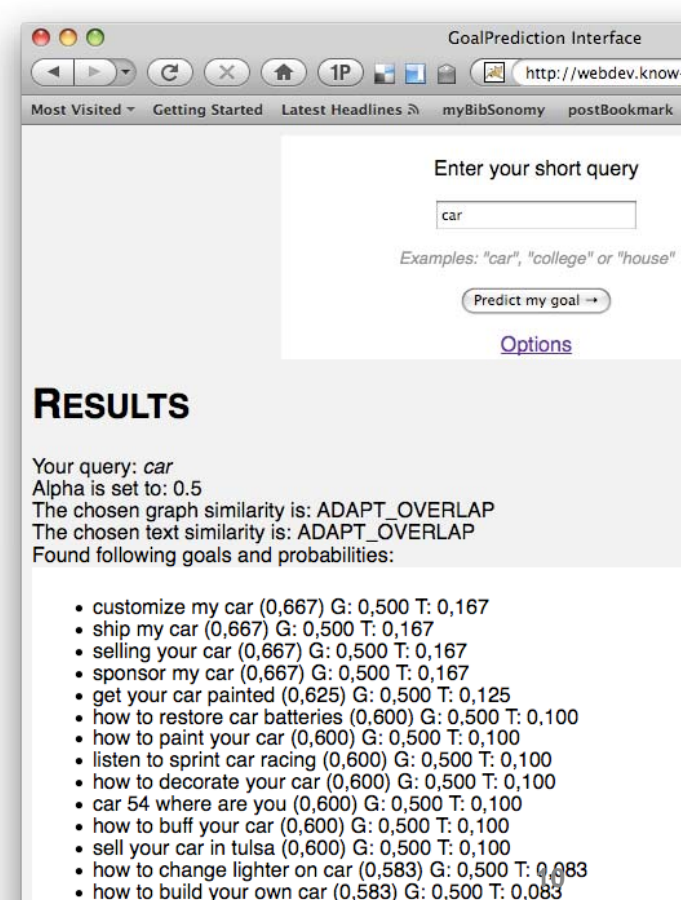
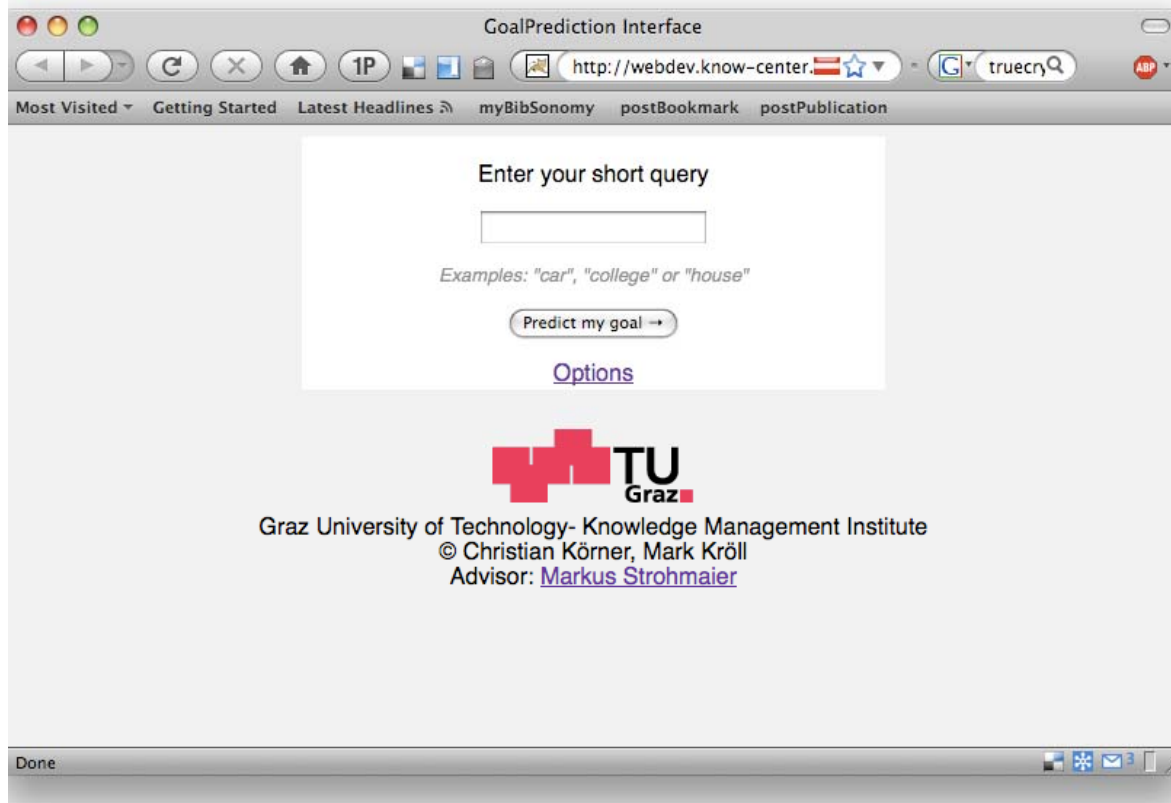
### Evaluation of a method to improve semantics in a folksonomy

- Example for quantitative evaluation

# Case Study 1: Goal Prediction Interface

Predicts a user's goal based on an issued search query

uses search query log information



# Evaluating the Goal Prediction Interface /

## 1

Three configurations with different parameter settings were selected for testing

### Preprocessing:

- a set of 35 short queries was drawn from the AOL search query log
- unreasonable queries were removed (e.g. “titlesourceinc”)
- Test participants were from Austria, therefore queries like “circuit city” and other brands were removed

# Evaluating the Goal Prediction Interface / 2

System received the 35 queries as input

For each of the queries the top 10 resulting goals were collected

Query: playground mat

buy playground equipment  
how to build a swing set  
covering dirt in a playground  
buy children plastic slides  
how to raise money for the playground  
how to weave a basket fifth grade project

# Evaluating the Goal Prediction Interface /

## 3

User had to classify the resulting goals into three classes

### Relevance Classes and Examples

- **Class 1:** *Plausible User Intention* - The proposed query represents a likely goal the user wants to achieve.  
Query: “anime”  
Intention: “how to draw manga” or “how to draw anime”
- **Class 2:** *Potential User Intention* - The proposed query represents an unlikely but possible user goal.  
Query: “Boston herald”  
Intention: “getting around Boston”
- **Class 3:** *Clear Misinterpretation or Mismatch:* The proposed query has no relation with the initial query.  
Query: “car” Intention: “improve loudspeaker system”

# Evaluating the Goal Prediction Interface /

## 4

Examples of the classification:

Query: playground mat	Classification	Note
buy playground equipment	1	
how to build a swing set	1	
covering dirt in a playground	1	
buy children plastic slides	2	unlikely
how to raise money for the playground	2	unlikely
how to weave a basket fifth grade project	3	mismatch

# Evaluating the Goal Prediction Interface /

## 5

5 annotators labeled the top 10 results for 35 queries which were produced by three different configurations

Test participants had to label the best result set

This way the best configuration should be identified

- However for this task the agreement between the participants had to be calculated

# Inter-Rater Agreement / 1

also known as Cohen's kappa

$$\kappa = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)},$$

Pr(a).... relative observed agreement among testers

Pr(e).... hypothetical probability of chance agreement

# Inter-Rater Agreement / 2

K	Interpretation
0.0 - 0.2	Slight agreement
0.21 - 0.4	Fair agreement
0.41 - 0.6	Moderate agreement
0.61 - 0.8	Substantial agreement
0.81 - 1.0	Almost perfect agreement

# Evaluating the Goal Prediction Interface / 6

Average  $\kappa = 0.67$

- indicating substantial agreement

Raters	$\kappa$
I - II	0.57
I - III	0.65
I - IV	0.62
I - V	0.60
II - III	0.66
II - IV	0.77
II - V	0.71
III - IV	0.73
III - V	0.66
IV - V	0.73

In 83 % of the cases configuration 3 was chosen for the best result set

Configuration 3 also had the best precision (percentage of relevant goals)

	Configuration 1	Configuration 2	Configuration 3
Precision	0.43	0.39	0.72

# Quantitative Measures

Needed:

“Ground Truth” or “Golden Standard” - verified knowledge from a trusted resource

and/or

Baseline for calculations (naive or other measures)

## Case Study 2: Semantics in Folksonomies

Based on user behavior we created a (sub-)folksonomy  
which produces better tag semantics (synonyms)

We showed that pragmatics influence semantics in  
folksonomies

# Evaluating semantic similarity / 1

In an experiment we used four different measures to extract the users which contribute more to the semantics in a folksonomy.

*Objective:* find the method which works best to identify synonyms of tags in a social tagging platform (del.icio.us).

## Evaluating semantic similarity / 2

For each measure:

- By using cosine similarity on the tag vectors we computed for each tag its most similar tag.

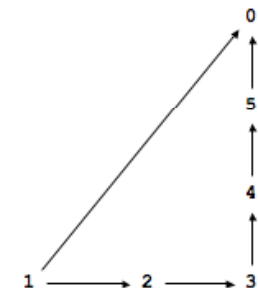
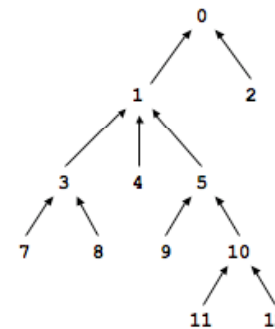
Question: How do we quantify the performance of each measure apart from anecdotal evidence?

# Evaluating semantic similarity / 3

Simple approach:

## WordNet distance

- length of path between two concepts which are contained within WordNet
- Disadvantages:
  - Does not take the structure of the network into account
  - Does not deal with multiple paths



# Evaluating semantic similarity / 4

Solution for these problems:

- Jiang-Conrath Distance

# Jiang-Conrath Distance

combines lexical taxonomy structure with the statistical information of a corpus

combined approach:

- edge-based (distance) approach
- node-based (information content) approach

[JiangConrath]

# Evaluating semantic similarity / 5

## What we did:

- For each tag we computed the most similar tag according to cosine similarity of the tag vectors produced by the four measures
- for each of these these tag pairs we computed the Jiang-Conrath distance (if both tags were present in WN)
- After that we calculated the average JCN distance of all mapped tag pairs and took this as an indicator for the semantic quality in our sub-folksonomy

# Evaluating semantic similarity / 6

## What we did cont.:

- As baselines we
  - selected random users from the folksonomy
  - used the complete folksonomy
- for both baselines the same procedures as described before were applied

## We showed:

- the best measure for the selection of the users
- not the complete folksonomy is needed for the emergence of the semantics within
- some users in a folksonomy generate “semantic noise” which does not facilitate the emergence of semantic structures in folksonomies
- tagging pragmatics influences semantics in folksonomies

# Limitations of this evaluation

only applies to words found in WordNet

- no slang or memes (“rick rolled”)
- no abbreviations (“UC LA”)

## Take home message(s)

Evaluation is a key factor for proofing that your work is correct

Good evaluation design takes time

Evaluation methods are manifold. There exists no absolute guide to evaluations.

Be creative!

Thank you for your attention!

# References

[Cohen] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1):37.

[Budanitsky] A. Budanitsky and G. Hirst, "Semantic Distance in Wordnet: An Experimental, Application-Oriented Evaluation of Five Measures," Proc. Workshop WordNet and Other Lexical Resources, Second Meeting of the North Am. Chapter of the Assoc. for Computational Linguistics, 2001.

[JiangConrath] J. J. Jiang and D. W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In Proc. of the International Conference on Research in Computational Linguistics (ROCLING). Taiwan, 1997.