is official
Implementation Partner of Intershop Communications AG

4FriendsOnly.com
Internet Technologies AG

a spin-off of

Fraunhofer
IDMT

SPONSORED BY THE
Federal Ministry
of Education
and Research

Automatic Taxonomy Extraction through Mining Social Networks

Mario Kubek, Jürgen Nützel, Frank Zimmerman
4FriendsOnly.com Internet Technologies AG (4FO AG)
MK(at)4FO.DE, JN(at)4FO.DE, FZ(at)4FO.DE

Outline

- Project GlobalMusic2one
- Social Networks as a Knowledge Resource
- Semantic Tag Relations in Social Networks
  - Asymmetric Tag Relations as a Basis to Calculate Tag Hierarchies
  - Tag Associations
- Taxonomy Extraction
  - Terminology and Basic Algorithm
  - Results and Enhancements
  - Fields of Application
- A Last.fm Taxonomy
- Conclusion
  - Summary
  - Further Research Tasks

The Project GlobalMusic2one [1/3]

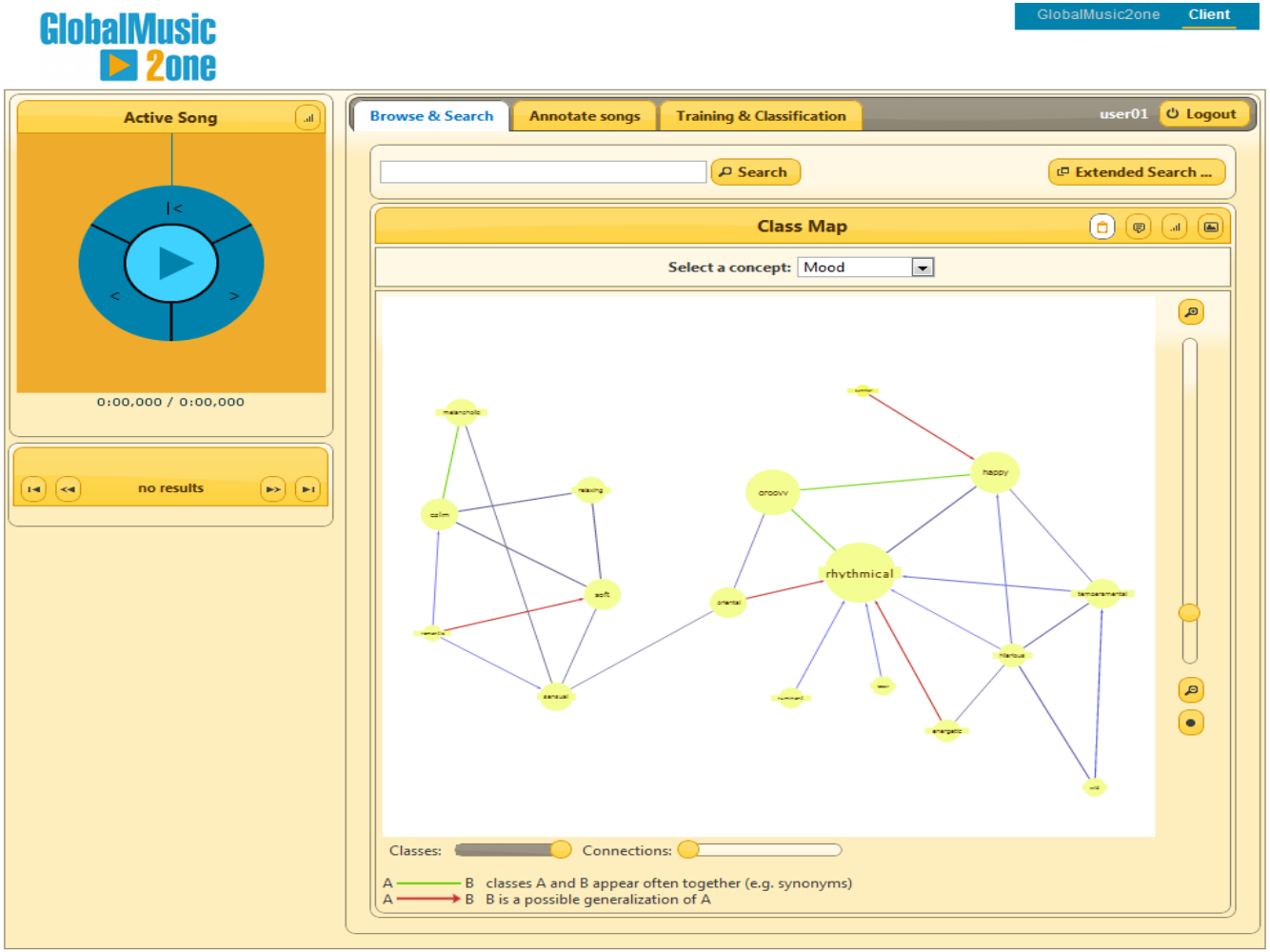
4FriendsOnly.com
Internet Technologies AG

- **GlobalMusic2one (GM2one)**
  - ***Project develops new adaptive methods for hybrid music search and recommendation of global music content***
  - ***Funded by the German Federal Ministry of Education and Research***
  - ***Participating members: Bach Technology GmbH, Fraunhofer Institute for Digital Media Technology (IDMT), Piranha Musik & IT AG and 4FriendsOnly.com Internet Technologies AG***
- **Objectives**
  - ***Creation of a software system that users train by adding new musical categories and assigning them example songs***
  - ***Identify musical qualities (e.g. genre, tempo, mood) by automatic content analyzation (Fraunhofer IDMT)***
  - ***System learns to recognize relationships of musical categories based on these training sets and musical qualities***
  - ***Visualization of the semantic relations of musical categories, called classes in GM2one***

Application in the Project GlobalMusic2one [2/3]

4FriendsOnly.com
Internet Technologies AG

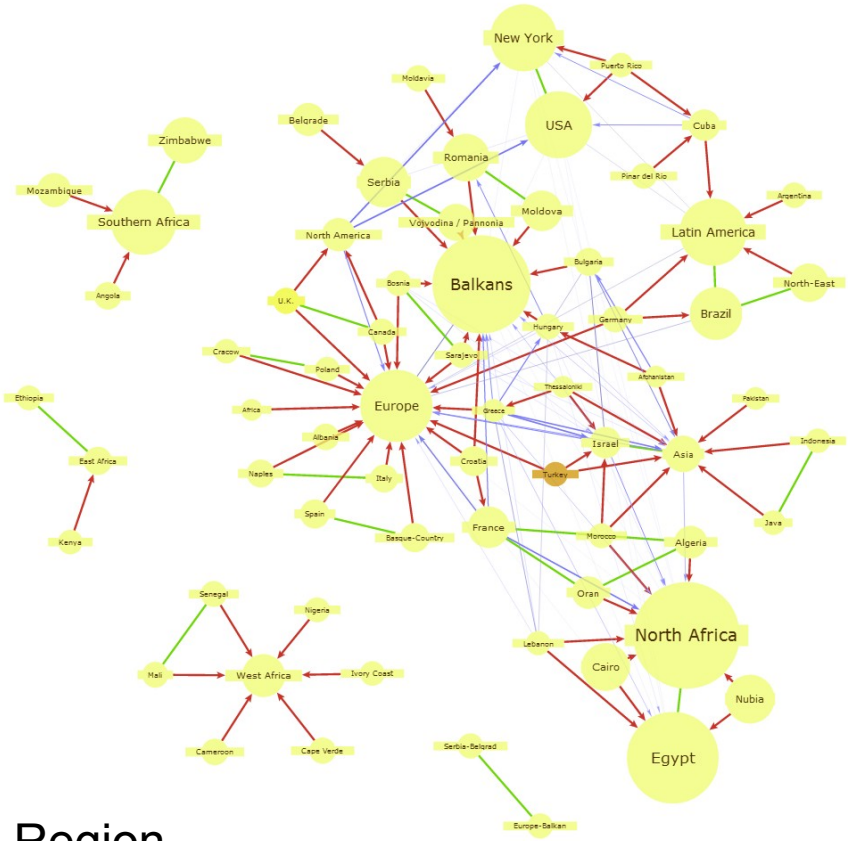
- **Prototype**



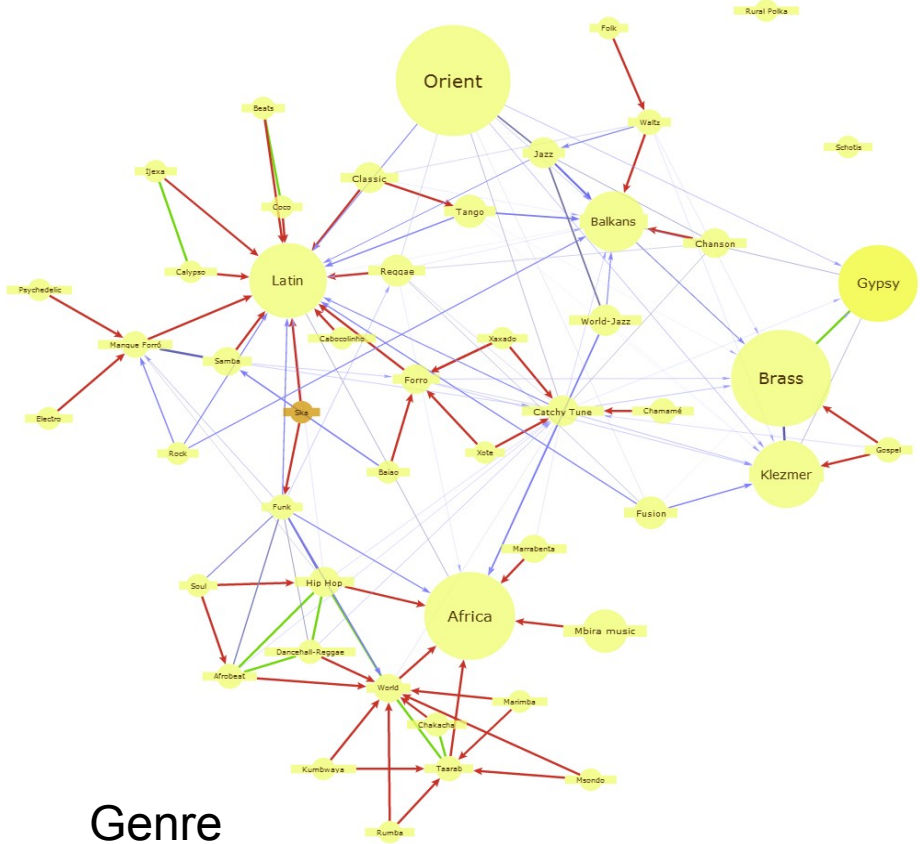
The Project GlobalMusic2one [3/3]

4FriendsOnly.com
Internet Technologies AG

- **GlobalMusic2one (GM2one)**
  - ***Class relations are calculated using the tag association formula***
- **Class Map Visualization in GM2one**



Region



Genre

Social Networks as a Knowledge Resource

- **Social networks are a very popular part of the Internet**
- **Users categorize items like pictures, videos and music with annotations, often called tags**
- **This collective categorization is often referred to as building a folksonomy**
- **Semantic relations among these tags can be determined using techniques from the natural language processing research**
- **They represent the basis to obtain further semantic dependencies between them e.g. to calculate tag hierarchies**

Semantic Tag Relations in Social Networks

- **Significant semantic tag relations can be detected via co-occurrence analysis when they often occur next to each other**
- **Prominent formulas for co-occurrence analysis like the poisson collocation measure and the log-likelihood ratio only yield symmetric tag relations**
- **Asymmetric tag relations however are a suitable basis to transform tag sets into semantic hierarchies or so-called taxonomies**
- **The conditional probability that a tag A occurs under the condition that it co-occurs with another tag B represents an asymmetric co-occurrence measure to acquire tag associations**

Asymmetric Tag Relations as a Basis to Calculate Tag Hierarchies

- **It can yield different values for the association of a tag A with a tag B and for the association of the same tag B with the same tag A**
- **When dealing with large datasets we can rely on conditional relative frequencies instead, so we just have to know the cardinality of tag sets**
- **The association of a tag A with a tag B can be determined by applying the following formula:**

$$Assn(A \rightarrow B) = \frac{|A \cap B|}{|A|}$$

- **With these asymmetric relations in form of tag associations it is possible to build up taxonomies of tags**

Semantic Tag Relations in Social Networks

Part of a tag cloud:

Item 1
Tag B
Tag A

Tag A
Tag B
Item 3

Tag A
Item 5

Item 2
Tag A

Tag A
Item 4

Item 6
Tag A

We count the tags and calculate asymmentric relations:

$|A| = 6$
$|B| = 2$
$|A \cap B| = 2$

$$\frac{|A \cap B|}{|A|} = 33\%$$

$$\frac{|A \cap B|}{|B|} = 100\%$$

Example: Items are cars. A = fast; B = BMW.

Results: 100% of BMW cars are fast. 33% of the fast cars are BMWs

Taxonomy Extraction[1/3]

- **Terminology:**
  - **Tag B is a subclass of a tag A, if the set of items tagged with B is a proper subset of the set of items tagged with A**
  - **Tag A is then a superclass of tag B**
  - **If tag B itself is a superclass of another tag C, then tag A is an indirect superclass of tag C**
  - **If a tag has more than one direct subclass, then these subclasses are called cohyponyms (they differ in at least one semantic feature)**

- **Basic algorithm (only calculation of subclasses demonstrated here, see paper for superclass calculation):**

Taxonomy Extraction [2/3]

1. **Calculate all tag associations of co-occurring tags using the previously shown formula**
2. **Determine direct subclasses of a tag A by applying the following rules:**
   - ***Find out all subclasses of tag A:***
     - *the association of another tag with tag A must be greater than the association of tag A with this other tag*
     - *the association of this other tag with tag A must be greater than 0.80*
     - *the association of tag A with this other tag must be less than 0.50 and greater than 0.09 to filter out less important tags*
   - ***Remove all indirect subclasses (subclasses of subclasses):***
     - *order all subclasses of tag A according to A's association with them*
     - *determine indirect superclasses of A by comparing all subclasses of tag A with each other and selecting those subclasses that have an association with another subclass of this set of at least 0.80*
     - *rule out these indirect superclasses*

Taxonomy Extraction[3/3]

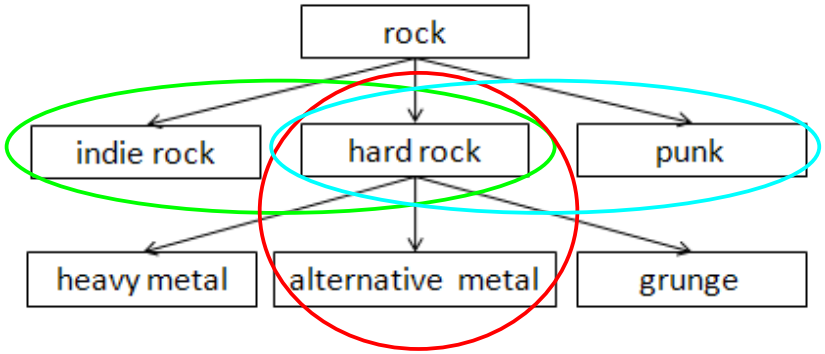
- **Result:**
  - *acyclic graph of hierarchically connected tags as nodes*
  - *works because people assign general and specific tags to items according to their knowledge*
- **Enhancements:**
  - *Clean up tag sets by removing stop words and rarely used tags*
  - *Apply methods of baseform reduction like stemming*
  - *Select only specific word forms like nouns through Part-Of-Speech tagging or dictionary lookups*
- **Fields of Application**
  - *New content recommendation algorithms*
  - *Automatically assign more general tags to content when it is tagged with specific tags only*
  - *Query Expansion with more general or more specific tags*
  - *Hierarchical tag cloud navigation*
  - *Detect inconsistencies among the annotations of a specific item*

A Last.fm Taxonomy [1/2]

- **We evaluated this algorithm by analyzing music tags in the social music service Last.fm**
- **1.2 million relations of 2800 tags of the 1500 most popular songs of 2009 in Germany have been calculated**
- **Tags that occurred less than 5 times were ruled out**

A Last.fm Taxonomy [2/2]

- **Small fraction of the entire taxonomy:**



Hard rock is 36% alternative metal
Alternative metal is 90% hard rock

Association: with of	rock	hard rock	indie rock	punk	heavy metal	alternative metal	grunge
rock		0.46	0.45	0.43	0.16	0.18	0.19
hard rock	1.0		0.36	0.53	0.32	0.36	0.34
indie rock	0.99	0.37		0.41	0.06	0.10	0.19
punk	0.98	0.56	0.01		0.16	0.16	0.28
heavy metal	0.95	0.87	0.17	0.41		0.62	0.26
alternative metal	0.96	0.90	0.26	0.38	0.58		0.22
grunge	1.0	0.83	0.45	0.64	0.23	0.22	

Conclusion

- **Summary**
  - *Introduced a method to calculate taxonomies from annotations in social networks based on statistical co-occurrence analysis*
  - *Discussed practical benefits of this approach*
  - *Example taxonomy of Last.fm tags presented*
  - *Outlined application of class associations in the project GM2one*
- **Further Research Tasks**
  - *As mentioned above cohyponyms differ in at least semantic feature. Is it always possible to technically confirm these presumed semantic differences?*
  - *Automatically assign tags from a taxonomy to content. A mapping from content features to the terms in the taxonomy is needed for this to work.*
  - *Calculate content similarities of sparsely or completely differently annotated items by matching their tags with entries in pre-calculated taxonomies and comparing their shared subtrees in the taxonomy*

Thank you for your attention!

Mario Kubek,
4FriendsOnly.com Internet Technologies AG (4FO AG)
MK(at)4FO.DE

Jürgen Nützel, CEO
4FriendsOnly.com Internet Technologies AG (4FO AG)
JN(at)4FO.DE

Frank Zimmermann
4FriendsOnly.com Internet Technologies AG (4FO AG)
FZ(at)4FO.DE

4FriendsOnly.com
Internet Technologies AG

Spin-off of

Fraunhofer
IDMT

SPONSORED BY THE
Federal Ministry
of Education
and Research

is official
Implementation Partner of Intershop Communications AG