GUIDING USER GROUPINGS
LEARNING AND COMBINING CLASSIFICATION
FOR ITEMSET STRUCTURING

Mathias Verbeke, Ilija Subašić and Bettina Berendt

MUSE
September 24, 2012, Bristol
MOTIVATION
MOTIVATION
MOTIVATION

Structuring is natural
MOTIVATION

Structuring is natural

... also for digital items
CONTRIBUTIONS

1. A new DM approach that learns an intensional model of user groupings and uses this to group new items.
   Identify structuring dynamics

2. New divergence measure

3. A study of grouping behaviour in a social bookmarking system
STRUCTURING
STRUCTURING
STRUCTURING

Art

CS

Music
STRUCTURING

Picasso

Monet

CS

Music
STRUCTURING

Helmets
The Roman Empire

Picasso
Monet

Music
CS
STRUCTURING

- Helmets
- The Roman Empire
- Picasso
- Monet
- CS
- Music
- Pop
- Rock
- Hip hop
- AI
...AND ITS DYNAMICS

• Goal: insight in structuring dynamics
dynamic conceptual clustering that simulates
the intellectual structuring process

• Two types of guides:
  1. own prior structuring
  2. structuring of peers
AT A GLANCE

Combination of 2 data mining tasks:

1. Learn model of structuring (classification)
   = intension: set of conditions for an object to belong to a certain class
   (vs. extension: list of objects in class)

2. Use intension or extension to structure new items
   A. based on own structuring
   B. based on k peers
GROUPING GUIDANCE
BASIC NOTATION

• **U**: the set of all users (used symbols: u, v, w)

• **T**: the set of all time points \{0, 1, ..., t_{max}\}, where t_{max} represents the time at which the last item arrives

• **D**: the set of all items (used symbol: d)

  \[ D_u^t \subseteq D: \text{the set of all } d \in D \text{ already considered by } u \in U \text{ at } t \in T \]

  \[ d_u^t \in (D \setminus D_u^t): \text{the item assigned to the structure by user } u \text{ at } t \]
GROUPING GUIDANCE

GROUPINGS AND CLASSIFIERS

- **G**: (machine-induced) groupings for each user’s items
- **C**: classifiers (i.e. intensions) learned for these groupings
  - **OG**: Observed Grouping
  - **GS**: Simulated Grouping, guided by self
  - **Gn**: Simulated Grouping, guided by \( n \) peers
INITIAL CLASSIFIER LEARNING

**Goal**: determine intensional definitions for the user-generated groupings

Each group is regarded as a class for which a definition needs to be calculated

Definitions used to assign new items to these groups
CLASSIFIER SELECTION

= selection of peer guides
CLASSIFIER SELECTION

= selection of peer guides

Requires divergence measure between groupings of non-identical item sets
CLASSIFIER SELECTION

= selection of peer guides

Requires *divergence measure* between groupings of non-identical item sets

... but existing measures require large overlap between sets
CLASSIFIER SELECTION

= selection of peer guides

Requires divergence measure between groupings of non-identical item sets

... but existing measures require large overlap between sets

Inter-guide measure of diversity:

\[
udiv(u, v) = \frac{1}{2} \left( \frac{1}{|G_u^t|} \sum_{x \in G_u^t} \min_{y \in G_v^t} gdiv(x, y) \right) + \frac{1}{|G_v^t|} \sum_{y \in G_v^t} \min_{x \in G_u^t} gdiv(y, x)
\]
CLASSIFICATION

Selected classifiers are used to classify the item under consideration

Two cases:

- **Self-guided classification**
- **Peer-guided classification**
SELF-GUIDED

$\text{OG}_u^0$: observed initial grouping, learned: intensional description via classifier learning from extension

$\text{OG}_u^{t_{\text{max}}}$ (observed)

$\text{G}20_u^{t_{\text{max}}}$ (simulated, 20 peers)

$\text{G}10_u^{t_{\text{max}}}$ (simulated, 10 peers)

$\text{G}5_u^{t_{\text{max}}}$ (simulated, 5 peers)

$\text{G}1_u^{t_{\text{max}}}$ (simulated, 1 peer)

$\text{GS}_u^0$ (simulated, self-guided)
SELF-GUIDED

$U$

$t_u^1$: item $a$ reaches $u$

$OG_u^0$: observed initial grouping, learned: intensional description via classifier learning from extension

apply classifier $OC_u^0$ to item $a$;
learn the new classifier $CS_u^1$

$OG_u^{t_{max}}$ (observed)

$G20_u^{t_{max}}$ (simulated, 20 peers)

$G10_u^{t_{max}}$ (simulated, 10 peers)

$G5_u^{t_{max}}$ (simulated, 5 peers)

$G1_u^{t_{max}}$ (simulated, 1 peer)

$GS_u^0$ (simulated, self-guided)
SELF-GUIDED

$OG_u^0$: observed initial grouping, learned: intensional description via classifier learning from extension

$t_u^1$: item a reaches u

$t_u^2$: item b reaches u

apply classifier $OC_u^0$ to item a;
learn the new classifier $CS_u^1$

apply classifier $CS_u^1$ to item b;
learn the new classifier $CS_u^2$

$OG_u^{t_{\text{max}}}$ (observed)

$G20_u^{t_{\text{max}}}$ (simulated, 20 peers)

$G10_u^{t_{\text{max}}}$ (simulated, 10 peers)

$G5_u^{t_{\text{max}}}$ (simulated, 5 peers)

$G1_u^{t_{\text{max}}}$ (simulated, 1 peer)

$GS_u^0$ (simulated, self-guided)
SELF-GUIDED

OG\textsubscript{u}\textsuperscript{0}: observed initial grouping, learned:
intensional description via classifier learning from extension

\( OG\textsubscript{u}\textsuperscript{0} \)

\(OG\textsubscript{u}\textsuperscript{0}\) (observed)

\( G\textsubscript{20}\textsubscript{u}\textsuperscript{t_{\text{max}}} \)
(simulated, 20 peers)

\( G\textsubscript{10}\textsubscript{u}\textsuperscript{t_{\text{max}}} \)
(simulated, 10 peers)

\( G\textsubscript{5}\textsubscript{u}\textsuperscript{t_{\text{max}}} \)
(simulated, 5 peers)

\( G\textsubscript{1}\textsubscript{u}\textsuperscript{t_{\text{max}}} \)
(simulated, 1 peer)

\( GS\textsubscript{u}\textsuperscript{0} \)
(simulated, self-guided)

\( t\textsubscript{u}\textsuperscript{1}\): item a reaches u

\( t\textsubscript{u}\textsuperscript{2}\): item b reaches u

\( t\textsubscript{u}\textsuperscript{3}\): item c reaches u

apply classifier \( OC\textsubscript{u}\textsuperscript{0} \) to item a;
learn the new classifier \( CS\textsubscript{u}\textsuperscript{1} \)
apply classifier \( CS\textsubscript{u}\textsuperscript{1} \) to item b;
learn the new classifier \( CS\textsubscript{u}\textsuperscript{2} \)
apply classifier \( CS\textsubscript{u}\textsuperscript{2} \) to item c;
learn the new classifier \( CS\textsubscript{u}\textsuperscript{3} \)
PEER-GUIDED

U:
- \(t_u^1\): item a reaches u
- \(t_u^2\): item b reaches u
- \(t_u^3\): item b reaches u

\(OG_u^{t_{\text{max}}}\) (observed) - 20 peers
- \(G20_u^{t_{\text{max}}}\) (simulated, 20 peers)
- \(G10_u^{t_{\text{max}}}\) (simulated, 10 peers)
- \(G5_u^{t_{\text{max}}}\) (simulated, 5 peers)
- \(G1_u^{t_{\text{max}}}\) (simulated, 1 peer)
- \(GS_u^0\) (simulated, self-guided)

W:
- \(t_w^1\): item e reaches w
- \(t_w^2\): item e reaches w
- \(t_w^3\): item f reaches w

\(OG_w^{t_{\text{max}}}\) (observed) - 20 peers
- \(G20_w^{t_{\text{max}}}\) (simulated, 20 peers)
- \(G10_w^{t_{\text{max}}}\) (simulated, 10 peers)
- \(G5_w^{t_{\text{max}}}\) (simulated, 5 peers)
- \(G1_w^{t_{\text{max}}}\) (simulated, 1 peer)

V:
- \(OG_v^{t_{\text{max}}}\): observed initial grouping, learned:
  intensional description via classifier learning from extension

\(OG_v^0\) (observed) - 20 peers
- \(G20_v^{t_{\text{max}}}\) (simulated, 20 peers)
- \(G10_v^{t_{\text{max}}}\) (simulated, 10 peers)
- \(G5_v^{t_{\text{max}}}\) (simulated, 5 peers)
- \(G1_v^{t_{\text{max}}}\) (simulated, 1 peer)
- \(GS_v^0\) (simulated, self-guided)
PEER-GUIDED

**U**
- $t_u^1$: item a reaches u
- $t_u^2$: item b reaches u
- $t_u^3$: item b reaches u

**V**
- $t_v^1$: item g reaches v

**W**
- $t_w^1$: item d reaches w
- $t_w^2$: item e reaches w
- $t_w^3$: item f reaches w

**OG**
- $OG_v^{t_{max}}$: observed initial grouping, learned:
  - intensional description via classifier learning from extension
  - apply classifier $OC_u^0$ to item g;
  - learn the new classifier $C1_v^1$

**GS**
- $GS_v^0$: (simulated, self-guided)

**G**
- $G1_u^{t_{max}}$: (simulated, 1 peer)
- $G5_u^{t_{max}}$: (simulated, 5 peers)
- $G10_u^{t_{max}}$: (simulated, 10 peers)
- $G20_u^{t_{max}}$: (simulated, 20 peers)

**OG**
- $OG_v^{t_{max}}$: (observed)

**G**
- $G1_v^{t_{max}}$: (simulated, 1 peer)
- $G5_v^{t_{max}}$: (simulated, 5 peers)
- $G10_v^{t_{max}}$: (simulated, 10 peers)
- $G20_v^{t_{max}}$: (simulated, 20 peers)
PEER-GUIDED

OG : observed initial grouping, learned:
intensional description via classifier learning from extension

OG : observed

OG max (observed)

G20 max (simulated, 20 peers)

G10 max (simulated, 10 peers)

G5 max (simulated, 5 peers)

G max (simulated, 1 peer)

GS 0 (simulated, self-guided)

V

t : item g reaches v

t : item h reaches v

G1 tmax (simulated, 1 peer)

G5 tmax (simulated, 5 peers)

G10 tmax (simulated, 10 peers)

G20 tmax (simulated, 20 peers)

OG tmax (observed)

W

t : item d reaches w

t : item e reaches w

t : item f reaches w

OG max (observed)

G20 max (simulated, 20 peers)

G10 max (simulated, 10 peers)

G5 max (simulated, 5 peers)

G1 max (simulated, 1 peer)

GS 0 (simulated, self-guided)

U

t : item a reaches u

t : item b reaches u

t : item b reaches u

OG max (observed)

G20 max (simulated, 20 peers)

G10 max (simulated, 10 peers)

G5 max (simulated, 5 peers)

G1 max (simulated, 1 peer)

GS 0 (simulated, self-guided)
PEER-GUIDED

**V**

1. $t_v^1$: item g reaches v
2. $t_v^2$: item h reaches v
3. $t_v^3$: item i reaches v

---

**U**

1. $t_u^1$: item a reaches u
2. $t_u^2$: item b reaches u
3. $t_u^3$: item b reaches u

---

**W**

1. $t_w^1$: item d reaches w
2. $t_w^2$: item e reaches w
3. $t_w^3$: item f reaches w

---

**OG_v^0**: observed initial grouping, learned: intensional description via classifier learning from extension

- **OG_u^tmax** (observed)
- **G20_u^tmax** (simulated, 20 peers)
- **G10_u^tmax** (simulated, 10 peers)
- **G5_u^tmax** (simulated, 5 peers)
- **G1_u^tmax** (simulated, 1 peer)
- **GS_u^0** (simulated, self-guided)

- **OG_w^tmax** (observed)
- **G20_w^tmax** (simulated, 20 peers)
- **G10_w^tmax** (simulated, 10 peers)
- **G5_w^tmax** (simulated, 5 peers)
- **G1_w^tmax** (simulated, 1 peer)
- **GS_w^0** (simulated, self-guided)

---

**apply classifier OC_u^0 to item g;**

**learn the new classifier C1_v^1**

**apply classifier CS_u^1 to item h;**

**learn the new classifier C1_v^2**

**apply classifier CS_w^2 to item i;**

**learn the new classifier C1_u^3**
DATASET

CiteULike dataset

sampled with p-core subgraphs to overcome sparsity

| # users | 377 |
| # documents | 11,400 |
| # tags | 12,982 |
| timeframe | 01/2009 - 02/2010 |
INITIAL GROUPING

Tagging as implicit structuring

- First 7 months to learn initial grouping
- Modularity clustering
INITIAL GROUPING

Tagging as implicit structuring

- First 7 months to learn initial grouping
- Modularity clustering

Initial classifier learning

- High dimensional input space (BoW of abstracts)
  Naive bayes
SIMULATING GROUPINGS

• Groups represented by language models

• Jensen-Shannon divergence as inter-group divergence

\[ JS(\Theta_x, \Theta_y) = \frac{1}{2} KL(\Theta_x, \Theta_z) + \frac{1}{2} KL(\Theta_y, \Theta_z) \]

• Normalized Mutual Information to compare groupings

\[ NMI(G, G') = H(G) + H(G') - \frac{H(G, G')}{\sqrt{H(G)H(G')}} \]
RESULTS
SIMILARITY DISTRIBUTION
CONCLUSIONS

• Investigate and simulate collaborative structuring
  Learning and combining classifiers for itemset structuring

• New divergence measure

• Tested on social-bookmarking platform for literature management
CONCLUSIONS

LIMITATIONS

• Observed groupings based on tag assignments

• Simple classifier

... but provides initial insights into grouping behaviour and behaviour of users in social bookmarking systems
FUTURE WORK

• **Applications**: (tag) recommendation and social search
  Adds new level to individual and social measures

• **Regrouping** based on peers

• **Hybrid measure** for itemset structuring
THANKS!

QUESTIONS?