Discovering user communities on the Web and beyond

Georgios Paliouras

Institute of Informatics and Telecommunications
National Center for Scientific Research “Demokritos”

e-mail: paliourg@iit.demokritos.gr
Web page: http://www.iit.demokritos.gr/~paliourg

Ubiquitous Knowledge Discovery for Users (UKDU),
Workshop at ECML/PKDD,
Berlin, 22 September 2006

Outline

Web problems and solutions

- Web $\equiv$ easy access to information and services.
- Problems: size, structure and dynamics of the Web.
- Tools to facilitate access:
  search engines, Web directories, portals, etc.
- They do not quite work.

Personalization

- Intelligent solutions: personalization, semantics, etc.
- Personalization requires knowledge about the users, i.e. user models.
- Can we build user models from recorded usage data?
- Respecting user privacy.

Our approach

- Focus on generic user models (stereotypes and communities).
- Off-line user modeling, on-line personalization.
- Early work: Personalize Web sites.
  - Model common user interests.
  - Identify patterns in user navigation.
- Current work: Personalize the Web.
  - Personalize Web directories.
  - Include semantics in navigation patterns.

Beyond the Web

- New opportunities:
  - Mobile access to the Web.
  - New types of device on the Internet.
  - New network types.
  - More content and new services.
- New problems:
  - Increased information overload.
  - More noise (e.g. spam).
  - New dangers.
- Our proposal: discovery of active user communities.

Constructing Stereotypes

[UM1999]

- Assume registered users.
- Users provide personal information, e.g. occupation, age, gender etc.
- Record usage of the site (Web page requests):
  `[ses301, usr15, sports.html, football.html, basketball.html, racing.html]`
- Web pages may be organized into categories:
  `SPORTS=[sports.html, football.html, basketball.html, racing.html]`

Constructing Stereotypes

- Target: Models that associate stereotypical behavior with personal characteristics, e.g.

```
IF age IN [20..30] AND gender=male
THEN [football.html, racing.html]
```

- Discovery method:
  1. Group pages into categories (unsupervised).
  2. Identify patterns in user behavior (unsupervised), e.g. `[football.html, racing.html]`
  3. Associate patterns with personal information (supervised).

Constructing Communities

[ECDL1998,SMC1999, AAAI2000, ICML2000, IwC2002]

- Problems with stereotypes:
  - Hard to acquire accurate personal information.
  - Privacy issues.
- Solution: Restrict models to patterns in user behavior.
- We call these user communities.
- Initial approach: cluster users/sessions.

Constructing Communities

Our Approach

- We are interested in behavior patterns rather than user clusters.
- Community models $\equiv$ clusters of pages.
- Such models can be used directly for personalization, e.g. recommendation.
- Essential to allow overlapping clusters.

Constructing Communities

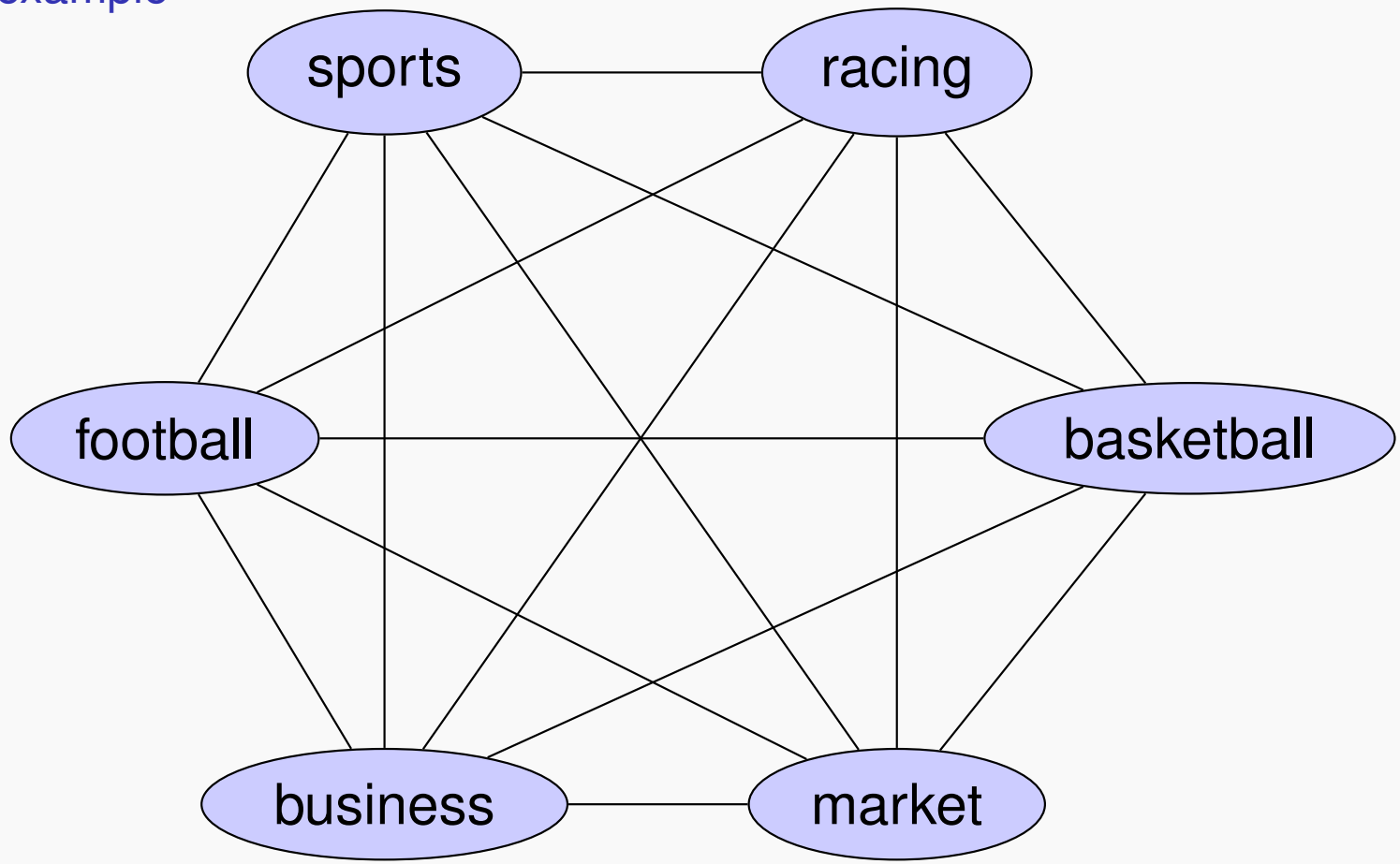
Graph-based clustering

Community models $\equiv$ cliques of Web pages:

1. Represent Web pages as bags of sessions:
   `[sports.html: ses1, ses12, ses123, ...]`
   `[racing.html: ses1, ses351, ...] ...`
2. Generate Graph $G =< E, V, W_e, W_v >$, where:
   V: pages, W_v freq. of occurrence,
   E: pairs of pages, W_e: freq. of co-occurrence.
3. Reduce graph connectivity by requiring $W'_e > T_c$, where $W'_e = W_e / \max(W_v^1, W_v^2)$.
4. Identify cliques in normalized G'.

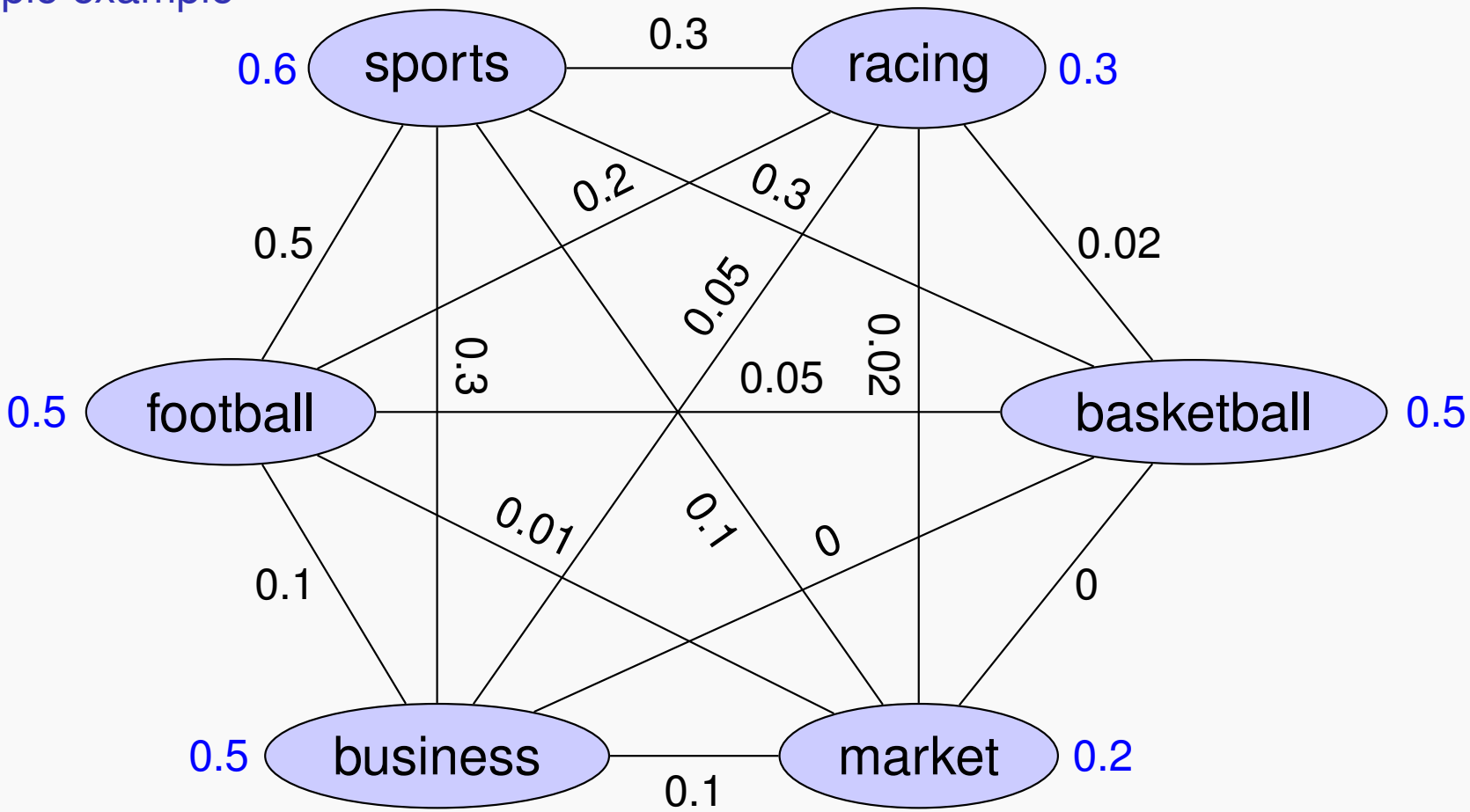
Constructing Communities

A simple example



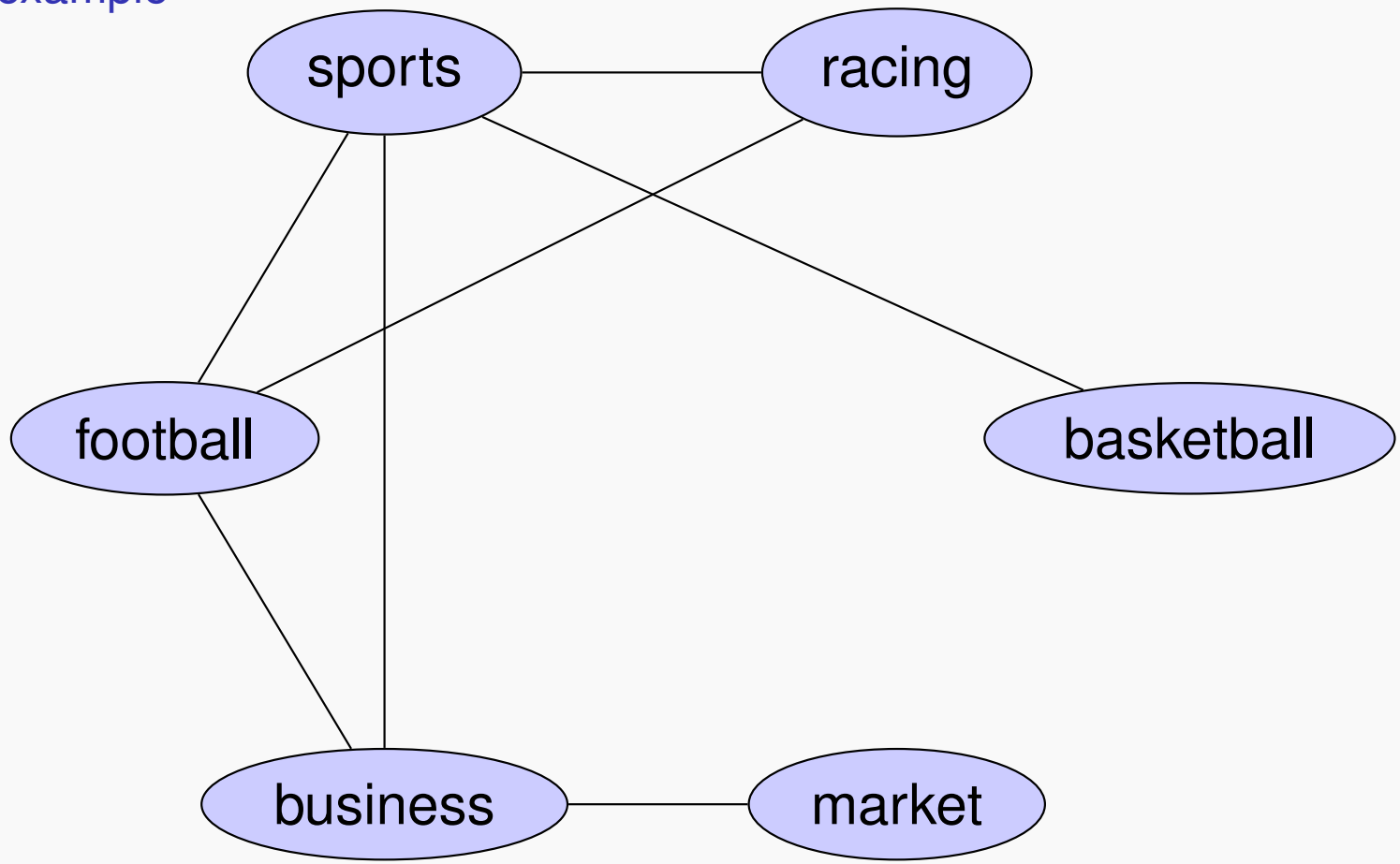
Constructing Communities

A simple example



Constructing Communities

A simple example



Modeling Web site navigation

- Model how users view the information.
- Initial approach: community models on page transitions, i.e. V is a set of page pairs in G,
  e.g. `[ses12, usr3, (sports.html, football.html), (football.html, racing.html)]`
- Interesting results, but may model discontinuous paths,
  `[(sports.html, football.html), (basketball.html, racing.html)]`
- Simplistic solution: remove discontinuous models.

Discovering grammatical models

[ICGI2004]

- Each Web page is a terminal symbol of a language L.
- Each user session is a string of the language.
- Assume strings are generated by an unknown grammar, modeled by a deterministic probabilistic SFA.
- Use grammatical inference to discover the automaton.

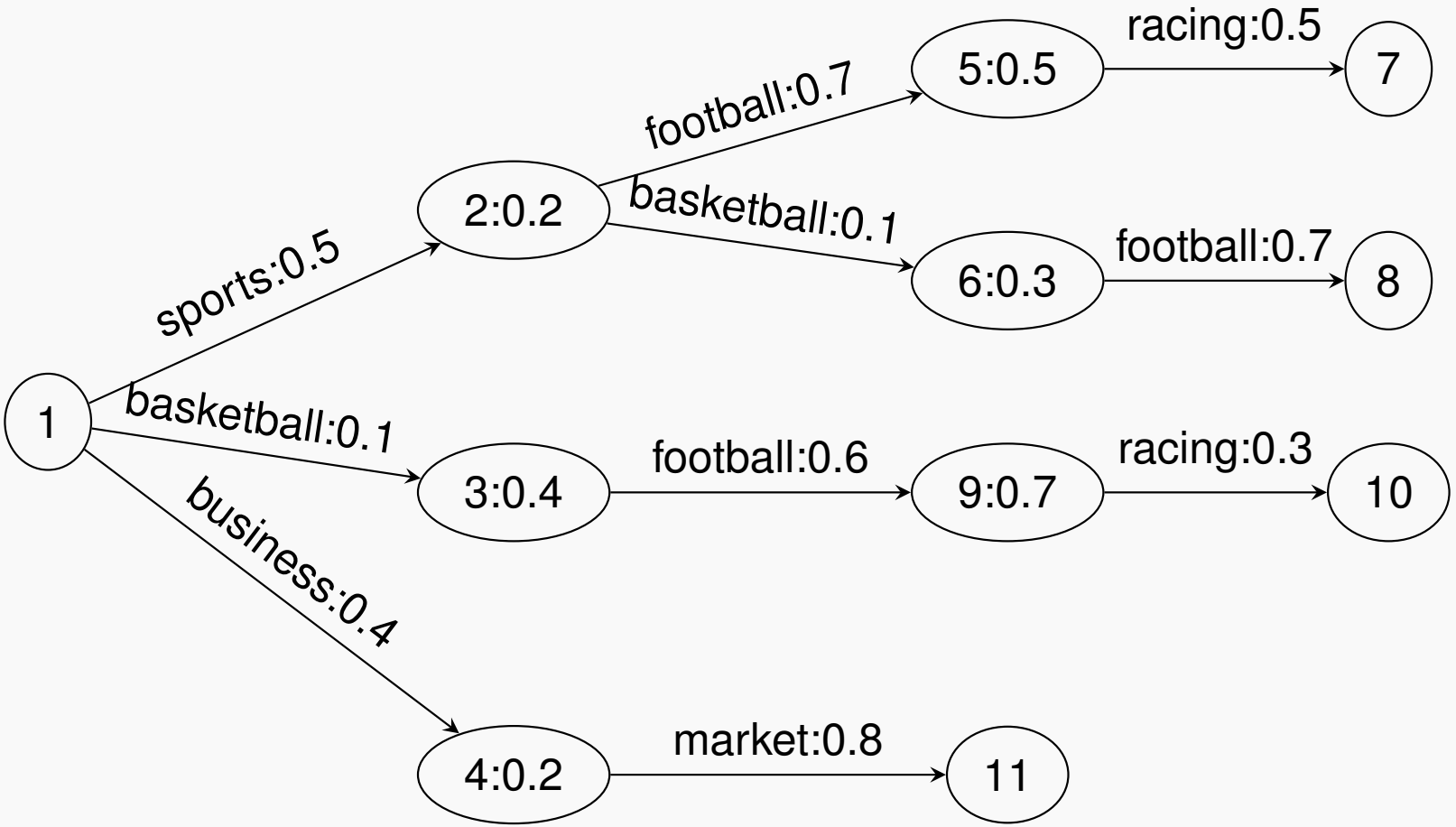
Discovering grammatical models

Grammatical inference

- Represent the data as a tree, in particular a PPTA: probabilistic prefix tree automaton.
- Iteratively merge compatible states, preserving determinism.
- Compatibility $\equiv$ similar outward transitions.
- Heuristic search of the space of compatible states.

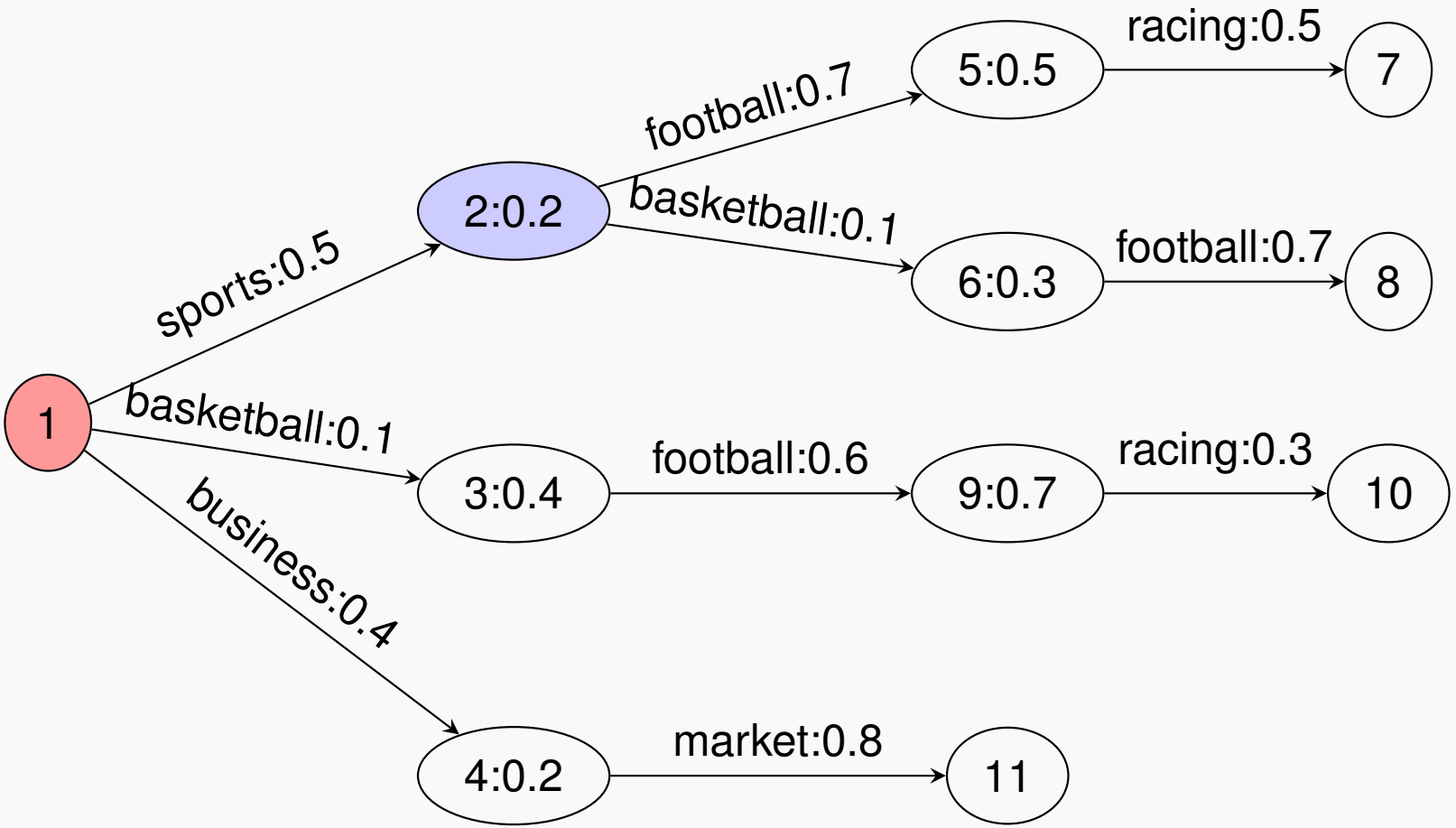
Discovering grammatical models

A simple example



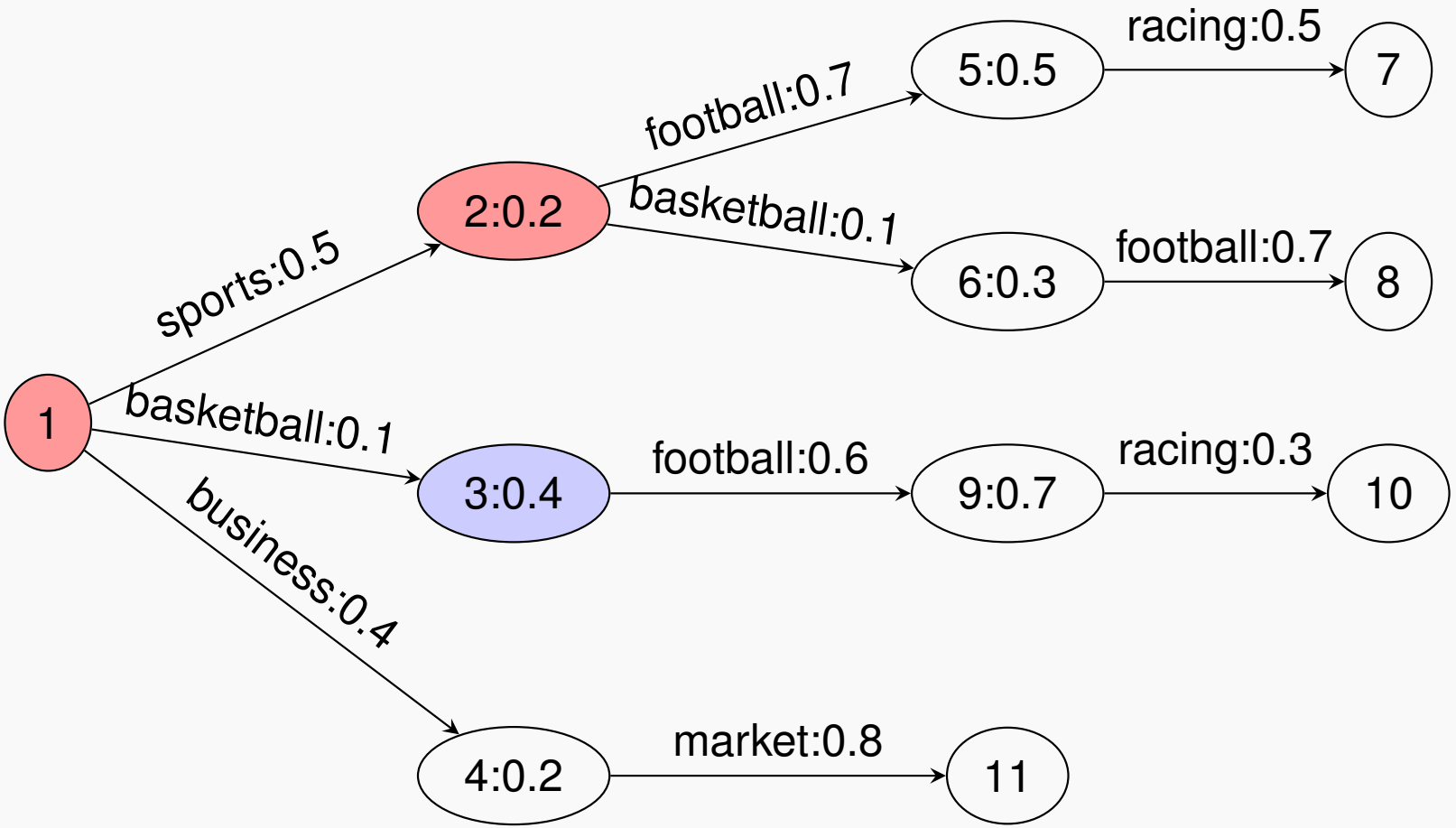
Discovering grammatical models

A simple example



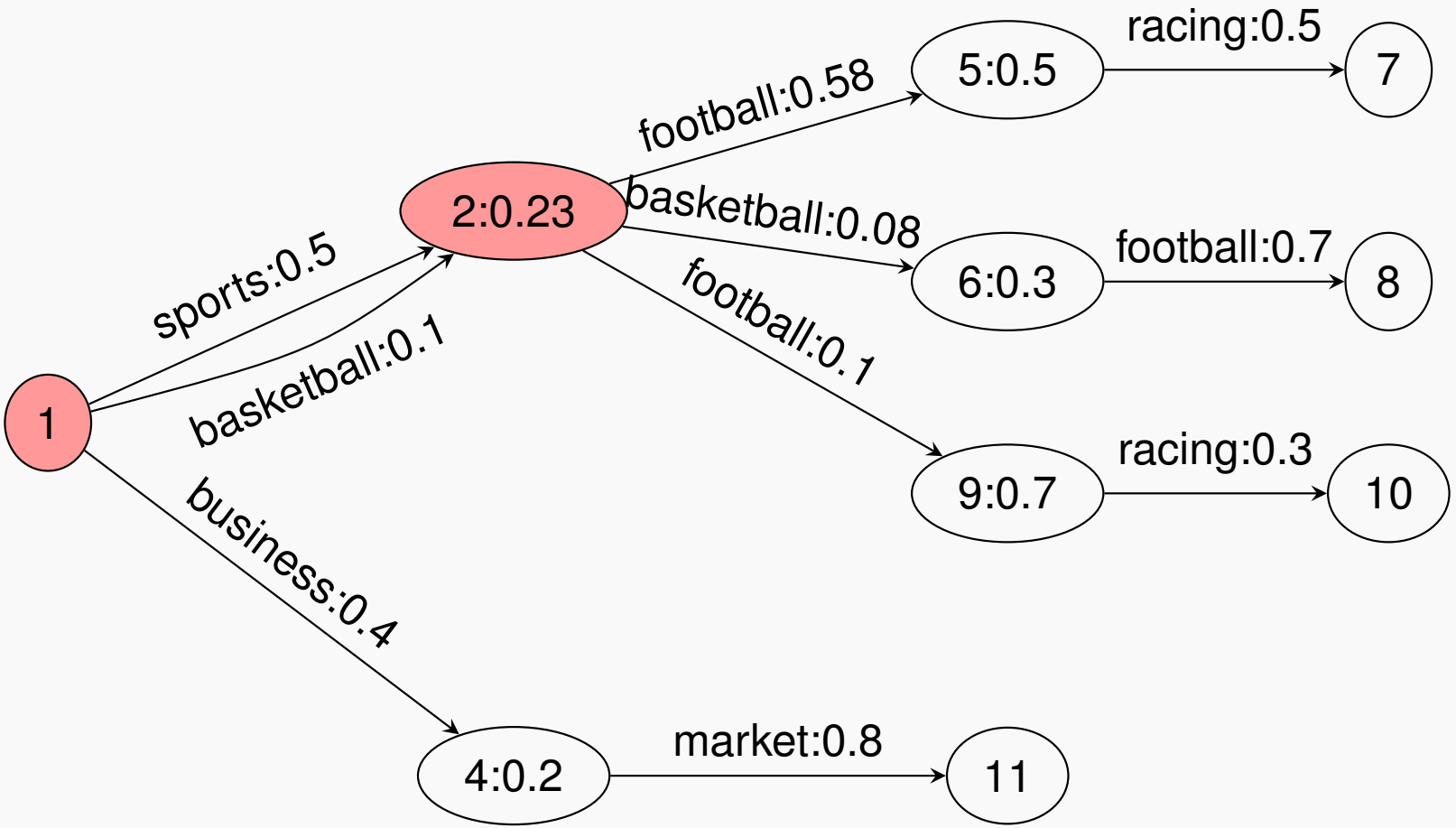
Discovering grammatical models

A simple example



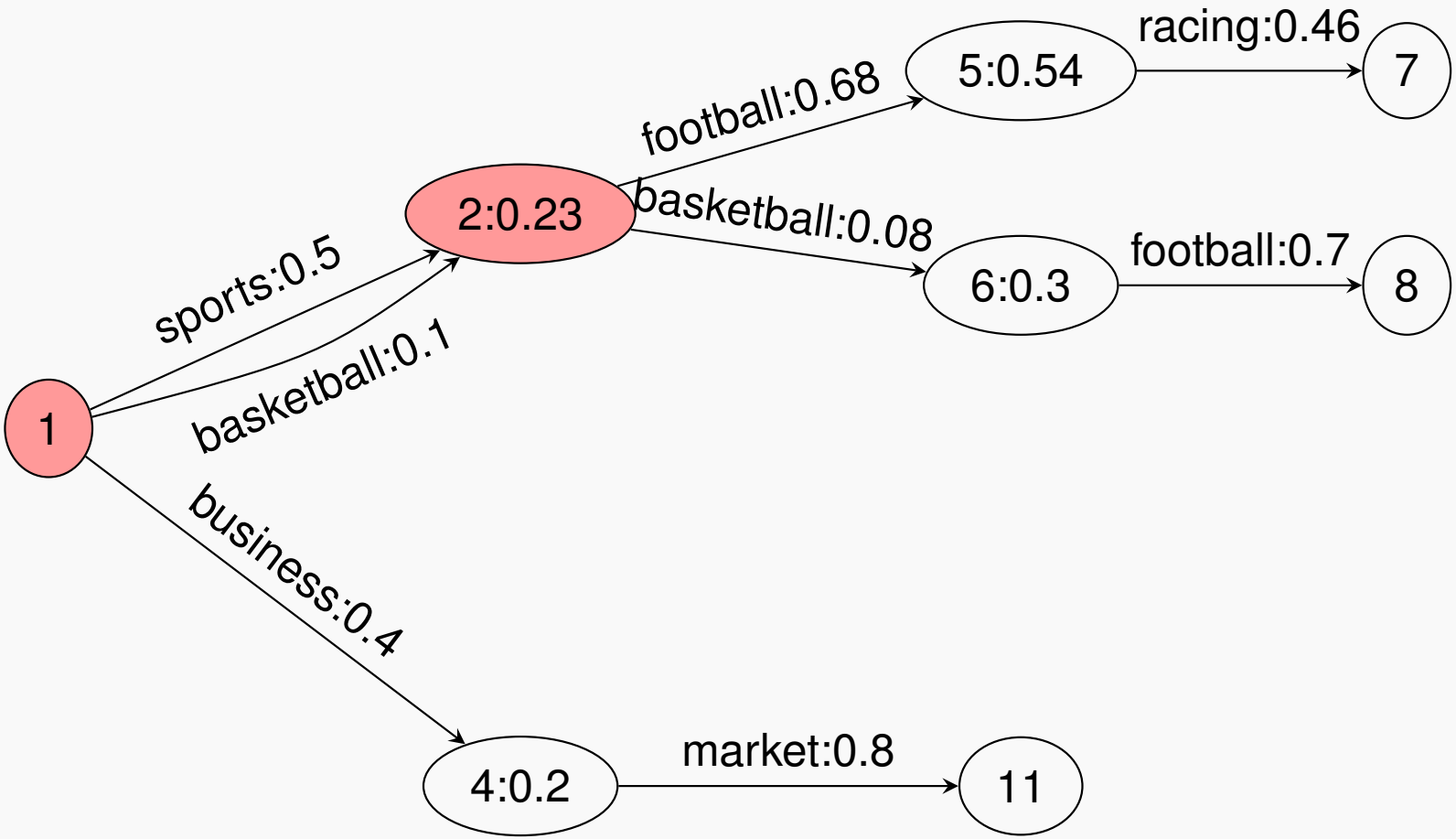
Discovering grammatical models

A simple example



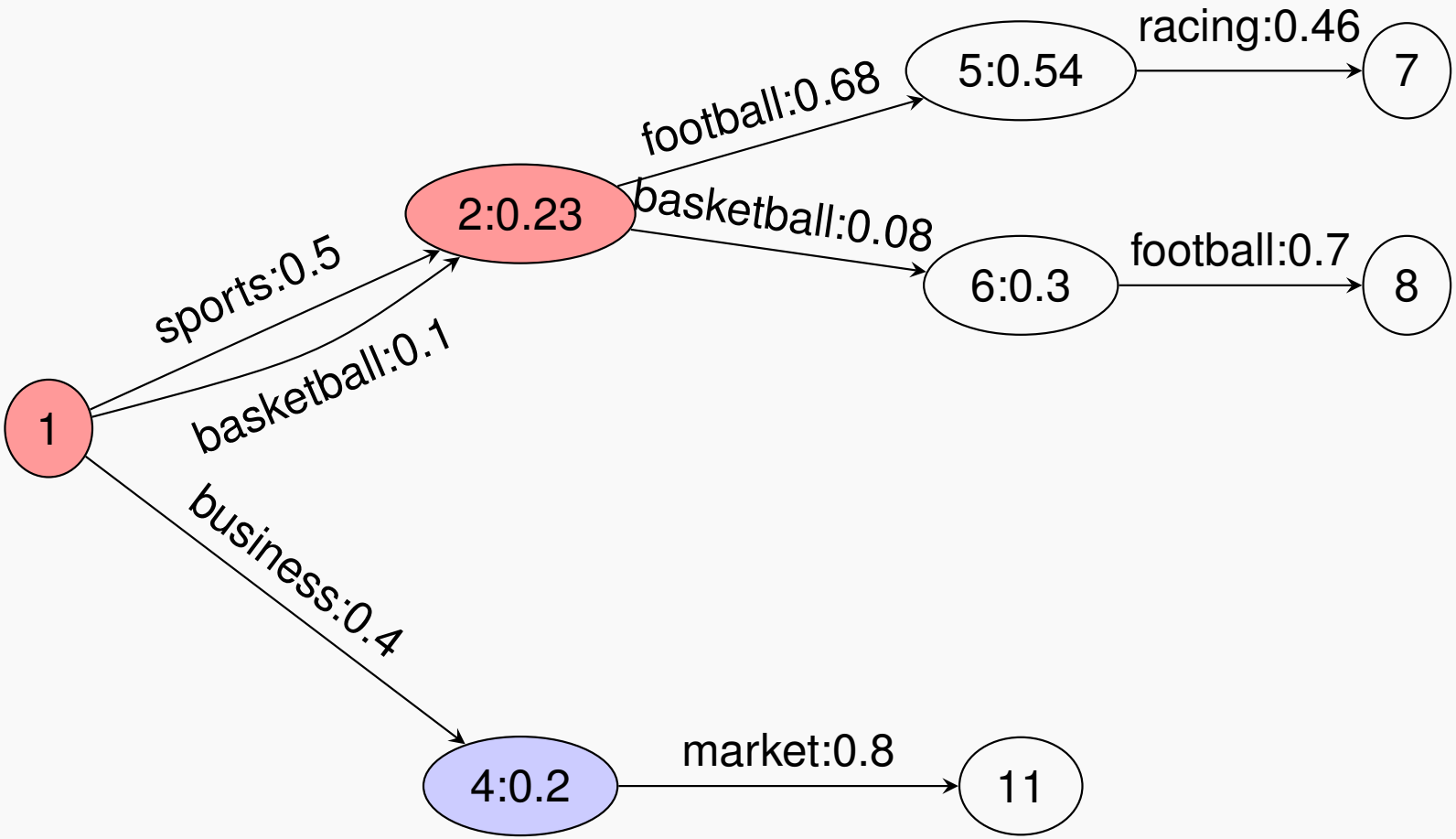
Discovering grammatical models

A simple example



Discovering grammatical models

A simple example



Discovering grammatical models

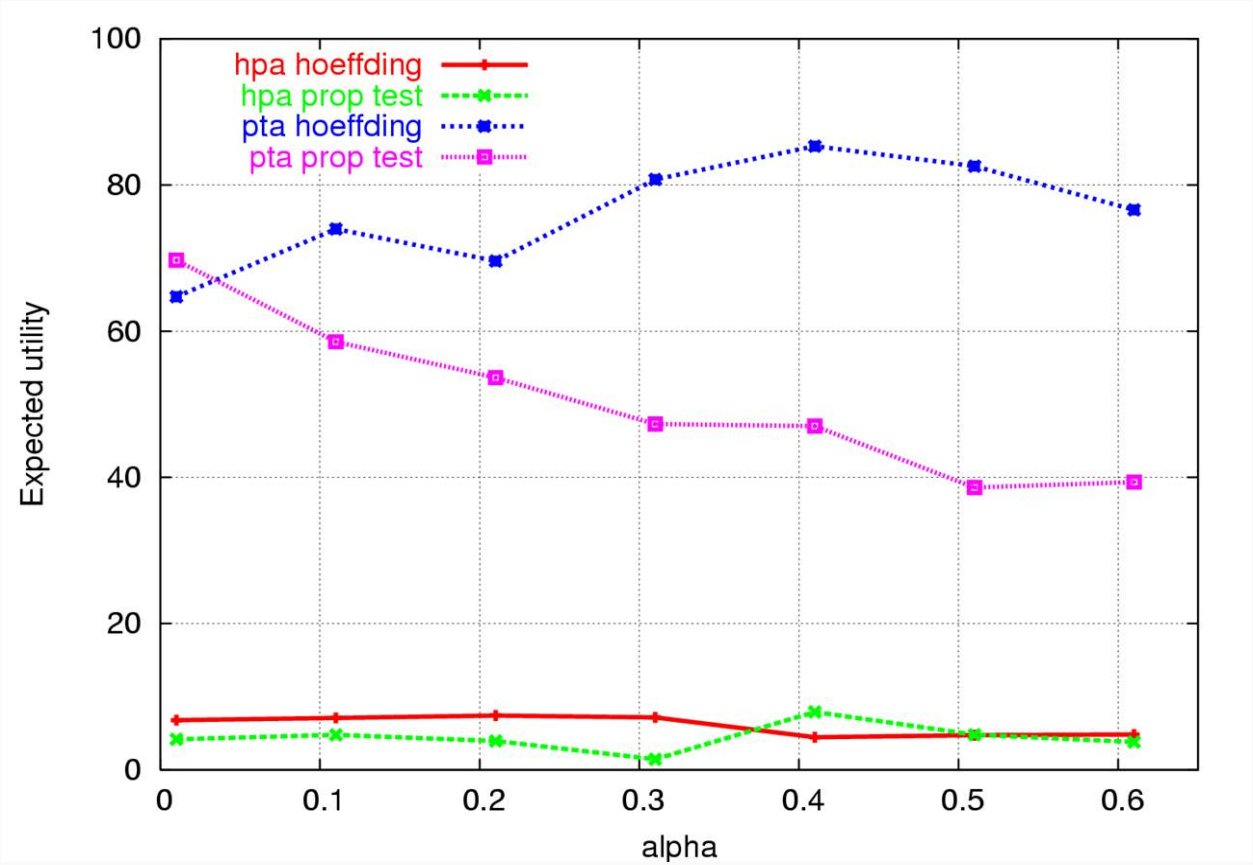
Experiments

- Recommendation on two large Web sites: MSWeb and a portal for chemistry.
- Evaluation process:
  1. Build model on part of the usage data.
  2. Hide the last page in the remaining sessions.
  3. Trace observed path on the automaton.
  4. Build recommendation list from current node's children.
- Evaluation measure (Expected Utility):

$$EU_a = \sum_{j=0}^{n-1} \frac{v_{aj}}{2^{j/h}}$$

Discovering grammatical models

Results



Modeling usage of the whole Web

The challenge

- The challenge of acquiring user models on the Web:
  - Usage data is voluminous.
  - Web structure is unknown and complex.
  - The users' interests, knowledge and behavior is diverse.
  - The thematic coverage of the data is very broad.

Community Web directories

[EWMF2003,HDMS2003,LNCS2004,UM2005]

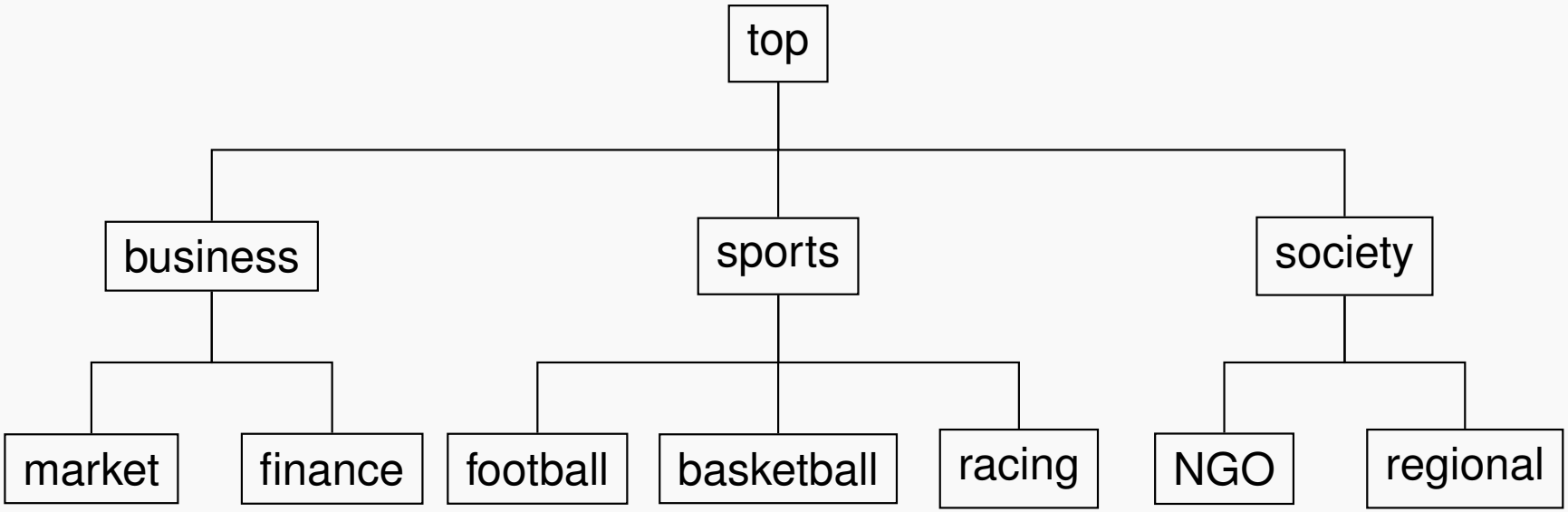
- Our approach: Combine modeling with Web directories
- A win-win scenario:
  - Web directories introduce thematic structure.
  - The size/dimensionality of the search space is reduced.
  - Directories are themselves in need of personalization.

Community Web directories

- Off-line user modeling:
  1. Map user sessions on the directory categories, i.e. each session becomes a small subdirectory.
  2. Create community Web directories.
  3. Prune non-representative branches.
  4. Remove redundant nodes, e.g. those without siblings.
- Personal Web directories constructed by assigning users to community directories and merging them.
- Personalized directories are small and provide quick access to interesting information.

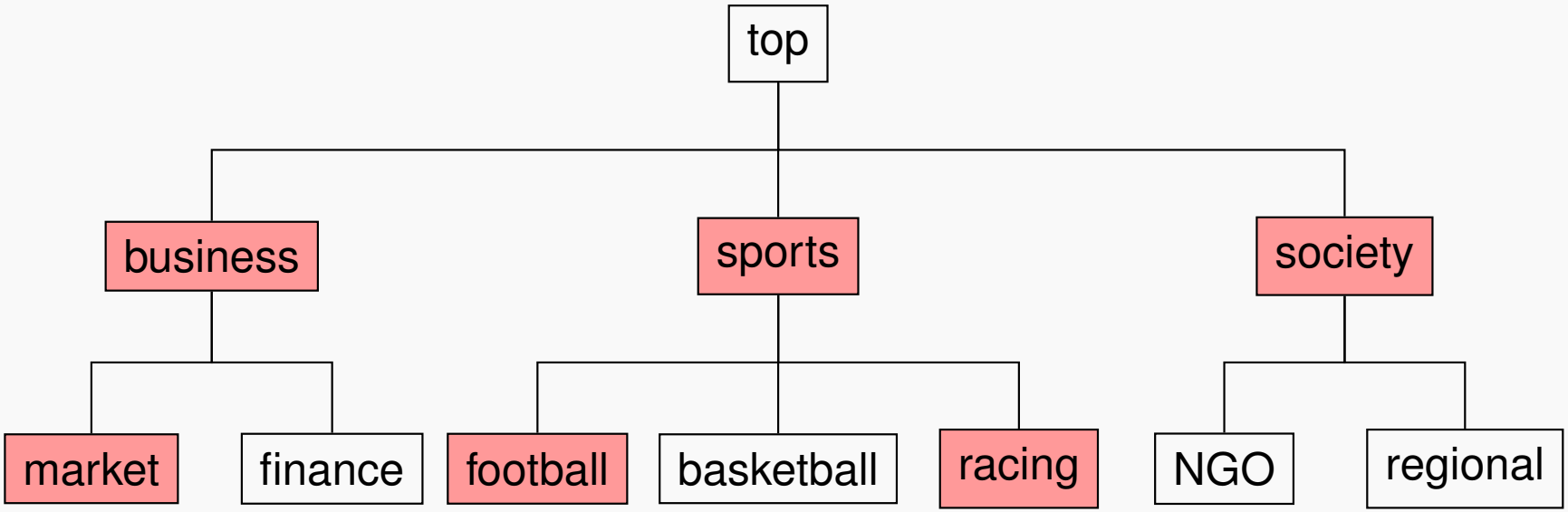
Community Web directories

A simple example



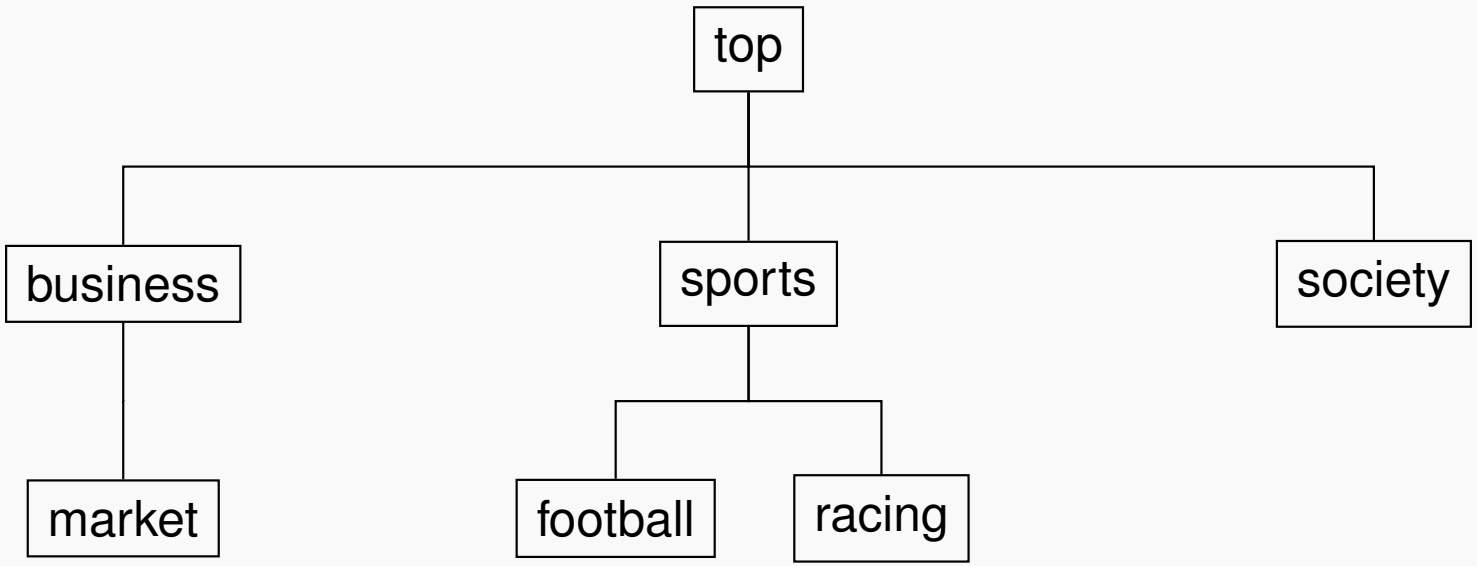
Community Web directories

A simple example



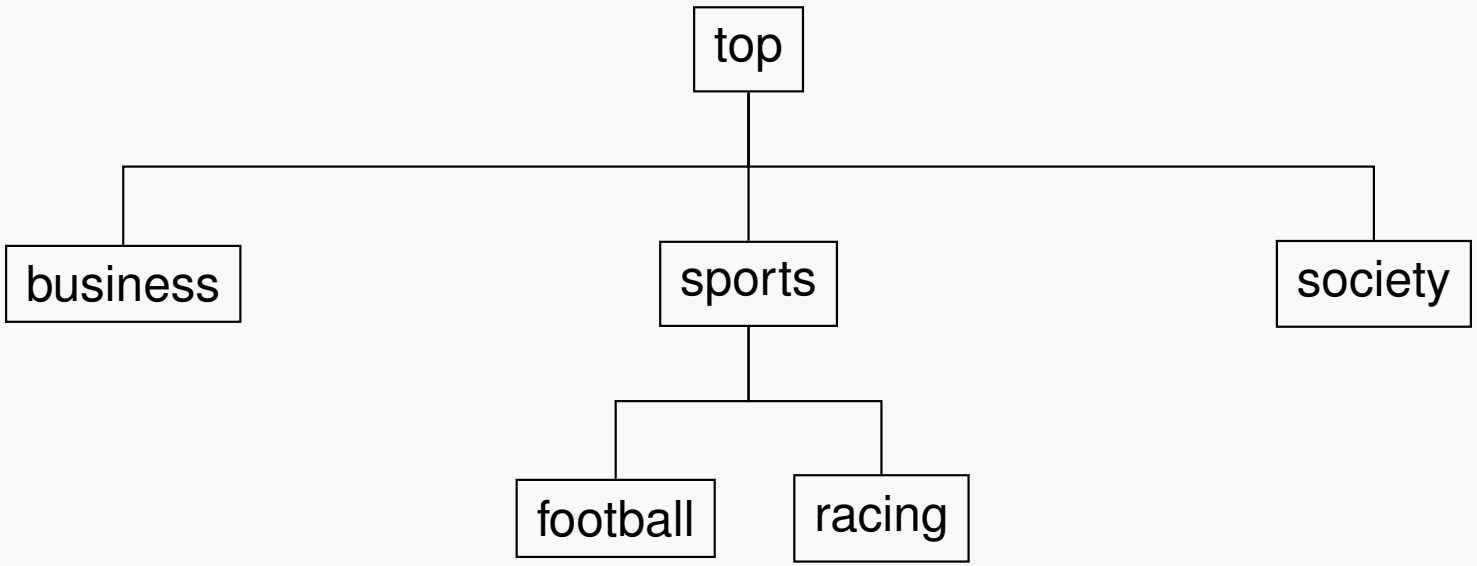
Community Web directories

A simple example



Community Web directories

A simple example



Community Web directories

Graph-based clustering

- A modified version of the method used for Web sites:
  1. Each directory category k_i becomes a node in the graph.
  2. Each page p_j is assigned a set K_j of categories, including all ancestors.
  3. For each occurrence of page p_j increase the weight of all $k_{ji} \in K_j$.
  4. For each co-occurrence of p_j and p_l increase the weight of all $(k_{ji}, k_{lm}), k_{ji} \in K_j, k_{lm} \in K_l$ edges.
  5. Reduce connectivity of the graph and find cliques.
  6. Construct a community directory for each clique.

Community Web directories

Latent-factor modeling

- Assume: a session u_i is due to a latent factor z_k, characterizing a community.
- Model the probability $P(u_i, c_j)$, where c_j a directory category:

$$P(u_i, c_j) = \sum_k P(z_k)P(u_i|z_k)P(c_j|z_k)$$

- Use Expectation Maximization to estimate the probabilities from the data.
- Construct a community directory for each factor, using the most representative categories: $P(c_j|z_k) > T_z$.

Community Web directories

Evaluation

- 781,069 records from ISP proxy server log.
- After cleaning and sessionization: 2,253 sessions
- Initial Web directory constructed with agglomerative document clustering (998 nodes).
- Repeated split of the data for modeling and evaluation.
- Hide last page from each evaluation session.
- Use observed pages to construct personal directory.

Community Web directories

Evaluation metrics

- *Coverage*: percentage of hidden pages covered by the personalized directories.
- *User Gain*:
  1. Position hidden page p_i in the directory.
  2. Measure *Click path*:

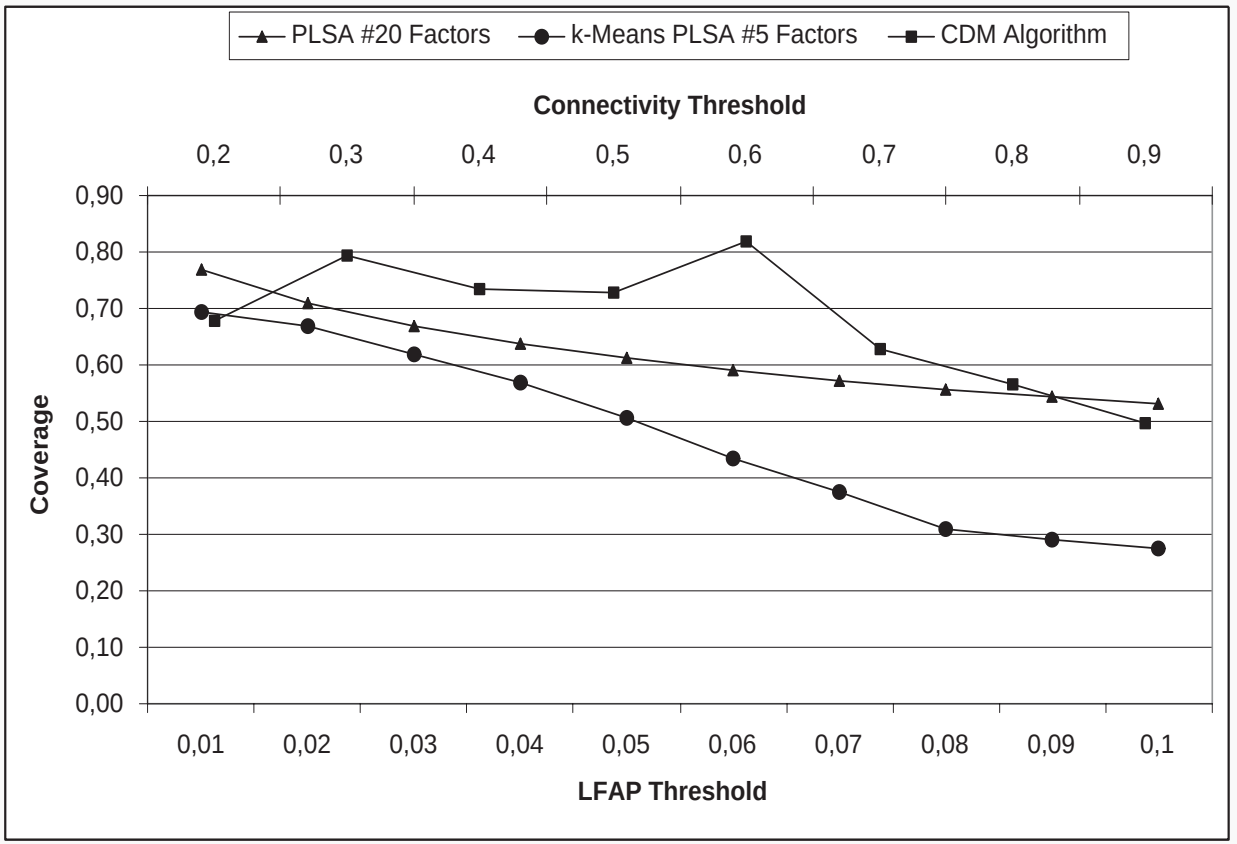
$$CP_i = \sum_{j}^{\texttt{depth}} j \times \texttt{branch\_factor}_j$$

  3. Measure average gain over original directory:

$$UG = \sum_{i} \frac{CP_i^{\texttt{gen}} - CP_i^{\texttt{pers}}}{CP_i^{\texttt{gen}}}$$

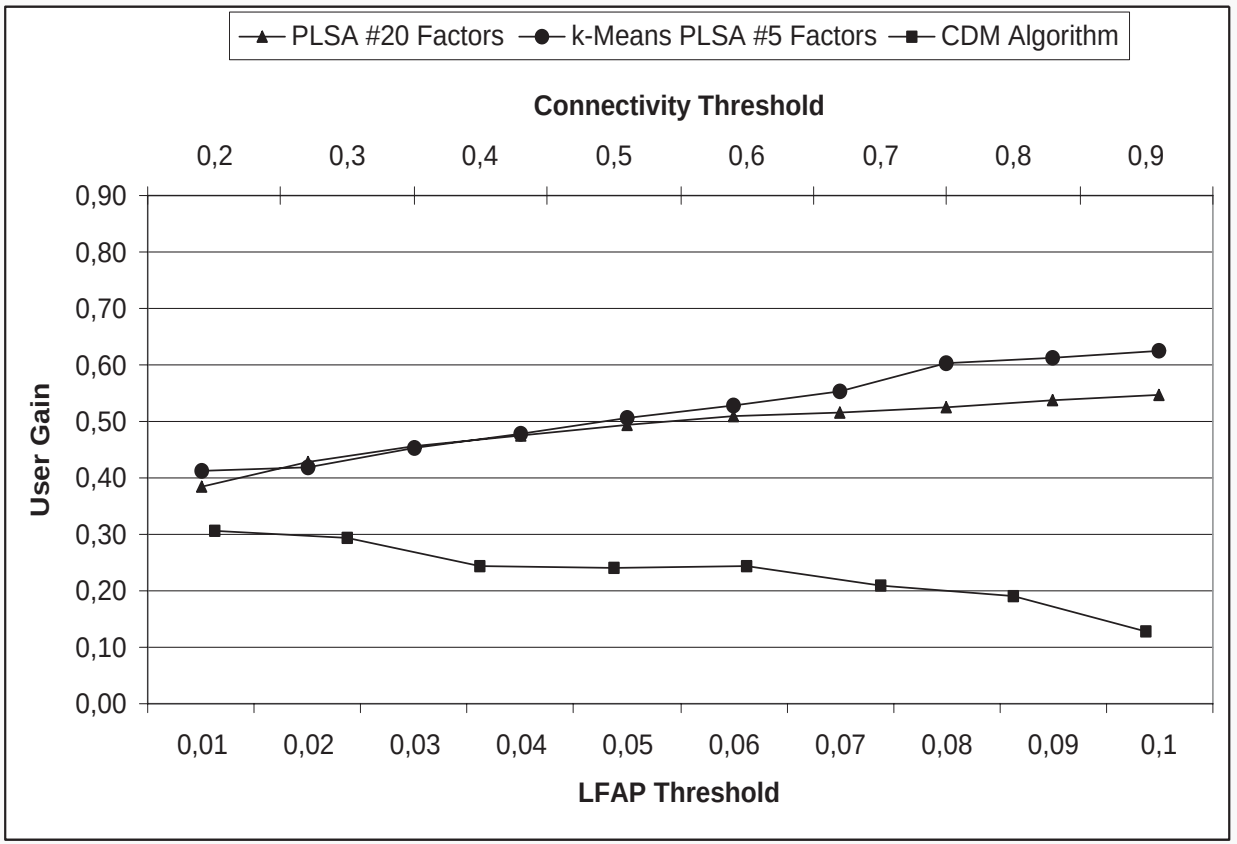
Community Web directories

Results



Community Web directories

Results



Modeling navigation on the Web

- Model how people navigate the Web.
- Acquire models from Web usage data, e.g. ISP.
- Can we apply the same methods as for a Web site?
- Statistics of Web page co-occurrence does not allow that.
- Our approach: model Web page similarity.

Content-Aware Navigation User Modeling with GI

[AAI:under review]

- Stick to grammars as navigation models.
- Key: each state is a cluster of the pages that lead to it.
- Each page (cluster) is represented as a word-frequency vector: `[goal=0.2,shot=0.1,basket=0,money=0.05]`.
- We can measure state compatibility by vector similarity, e.g. using the cosine metric.

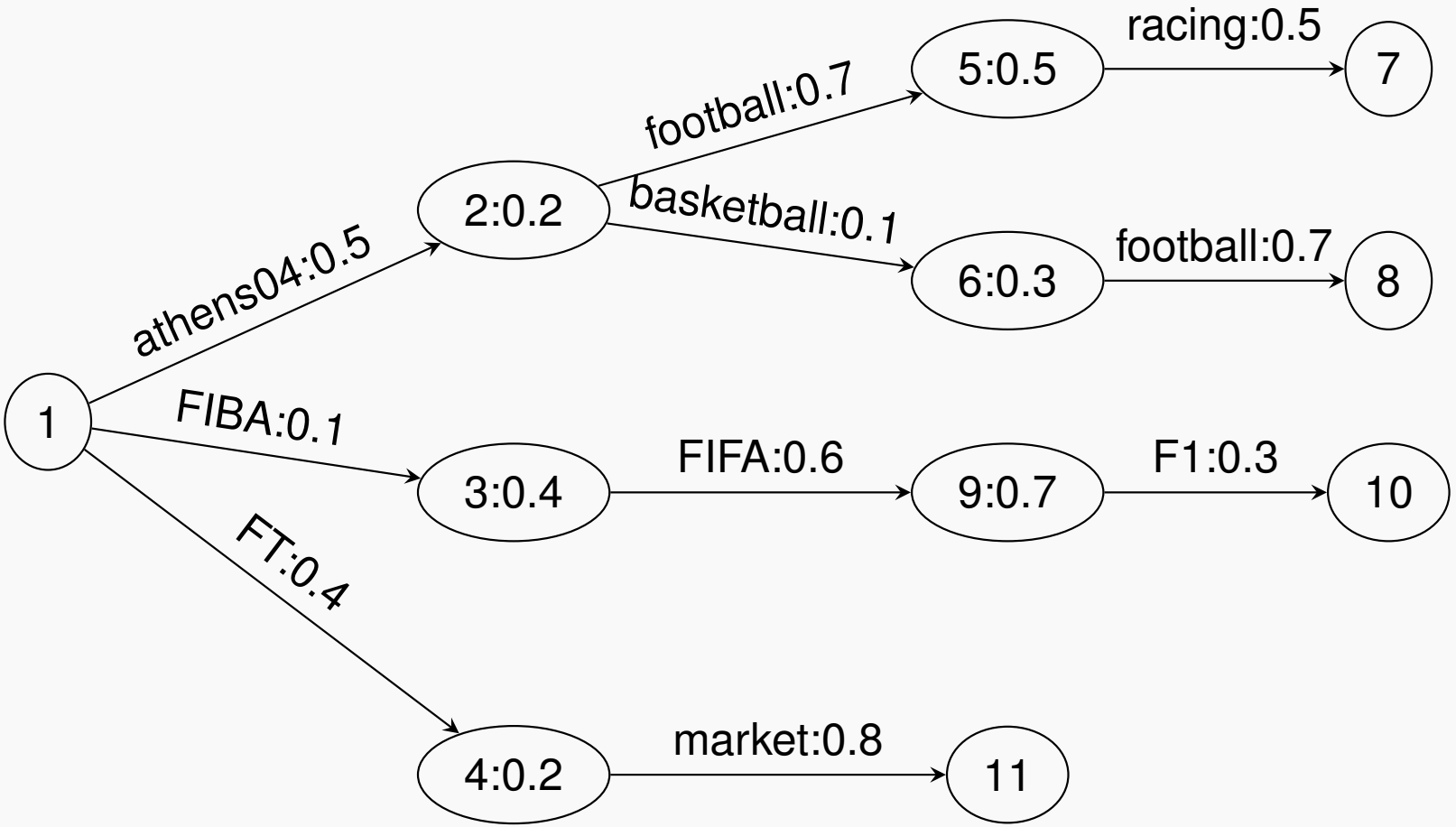
Content-Aware Navigation User Modeling with GI

Off-line modeling process

- Extend state compatibility to use content similarity:
  1. Measure usage and content similarity: $u(s_1, s_2)$, $c(s_1, s_2)$.
  2. Reject merge if $u(s_1, s_2) < T_u$ or $c(s_1, s_2) < T_s$.
  3. Normalize using the metric distributions in the PPTA.
  4. Combine by $\min$, $\max$, or weighted average.
  5. Merge most compatible pair of states.

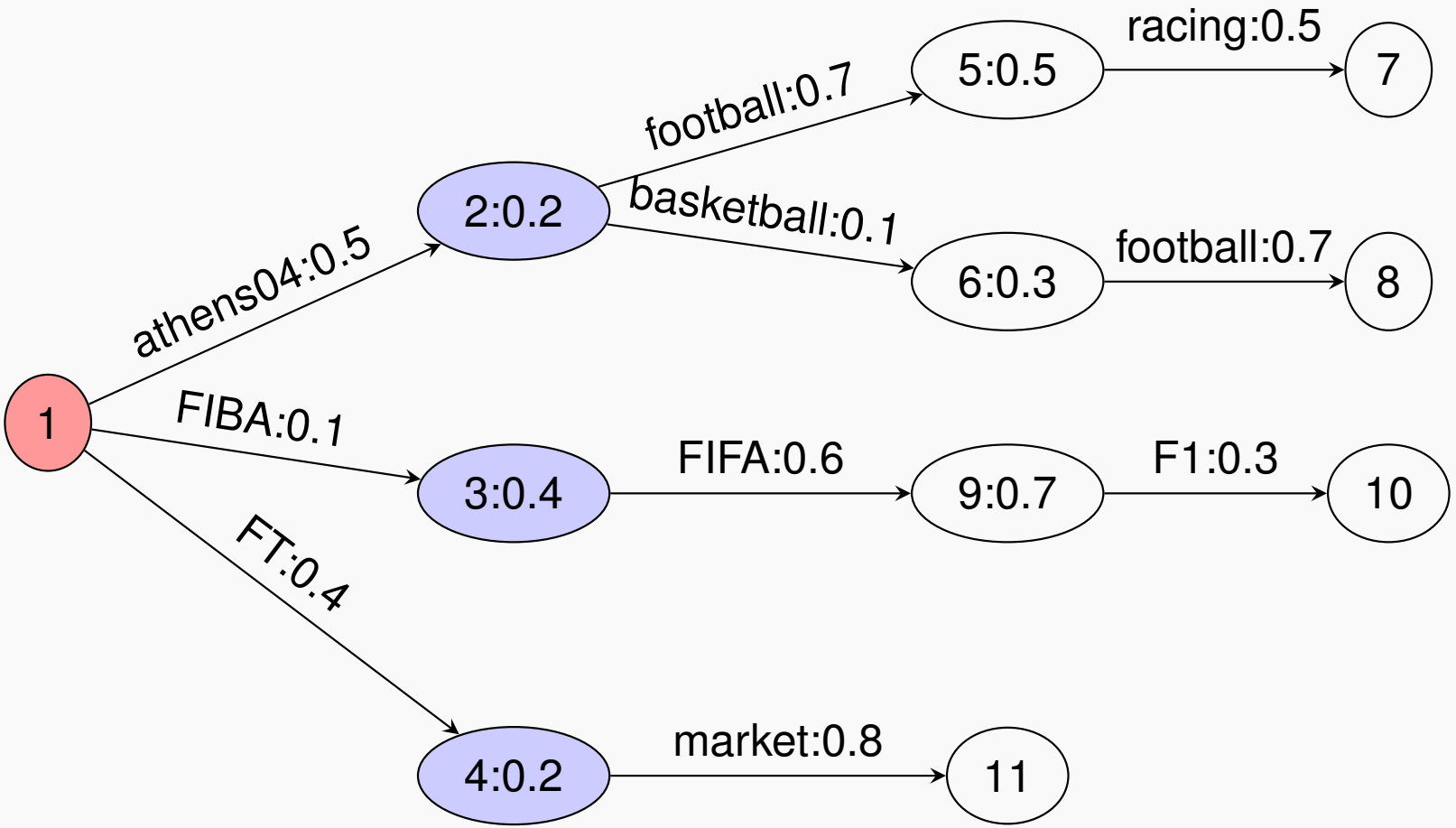
Content-Aware Navigation User Modeling with GI

A simple example



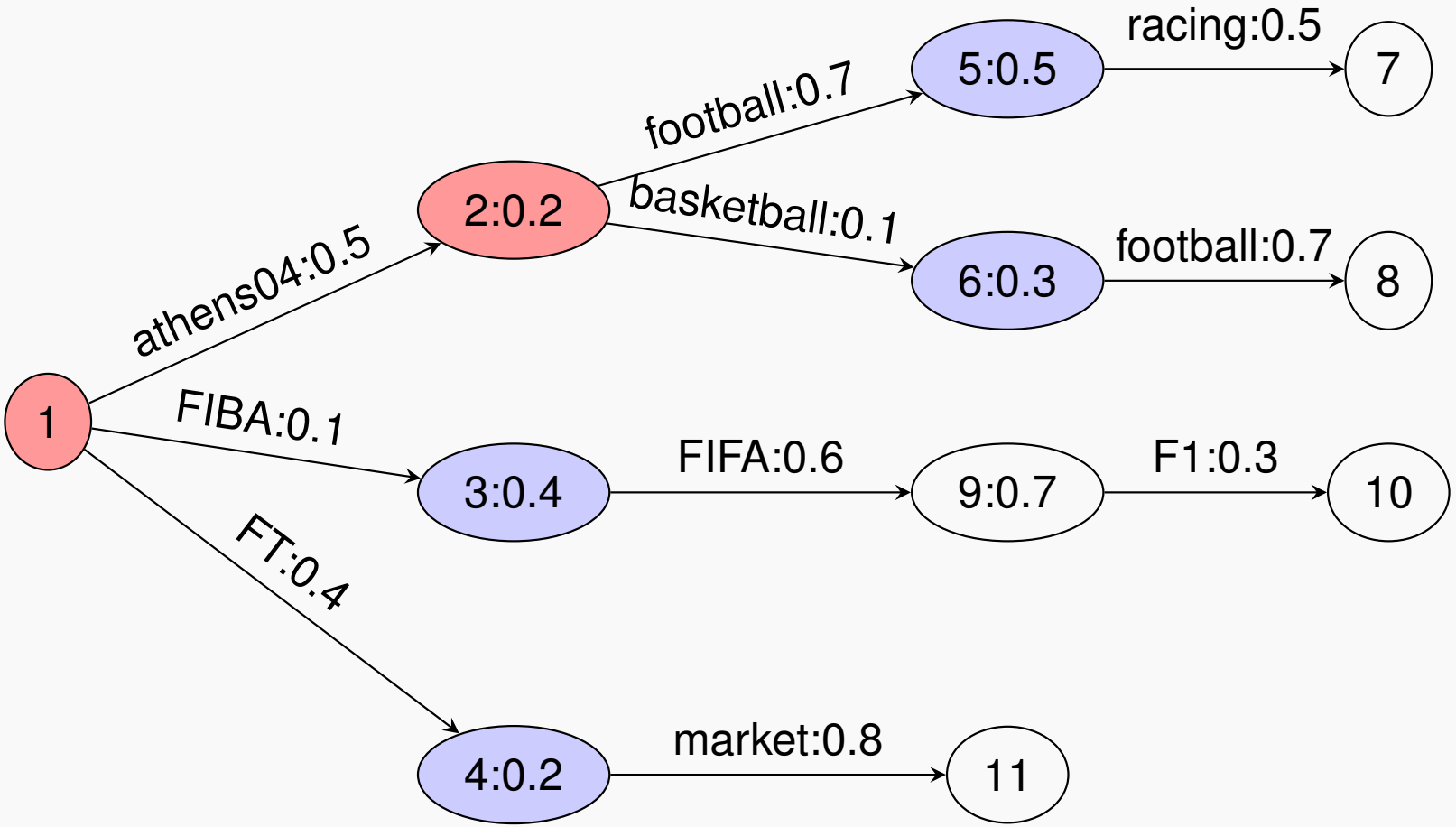
Content-Aware Navigation User Modeling with GI

A simple example



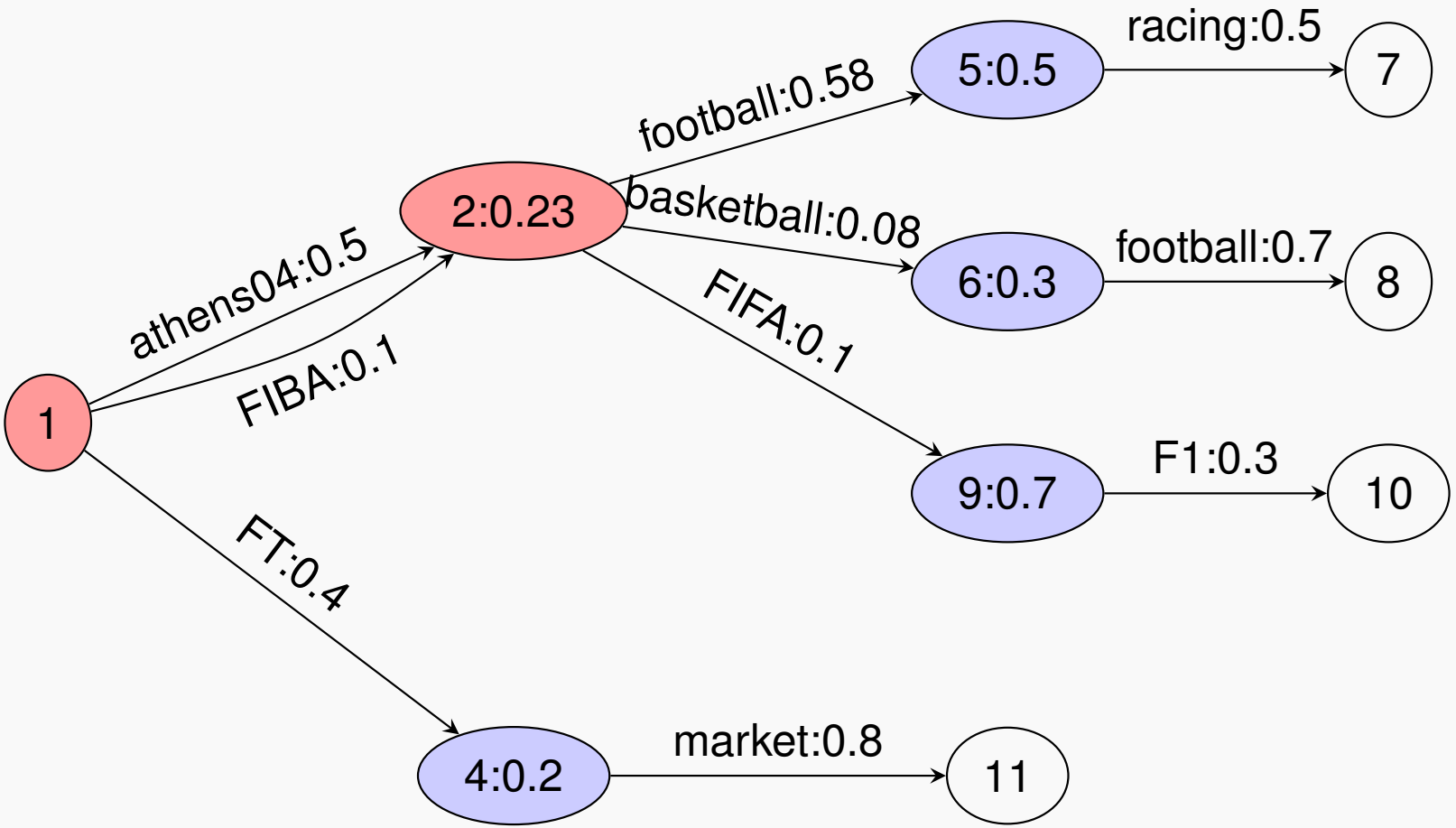
Content-Aware Navigation User Modeling with GI

A simple example



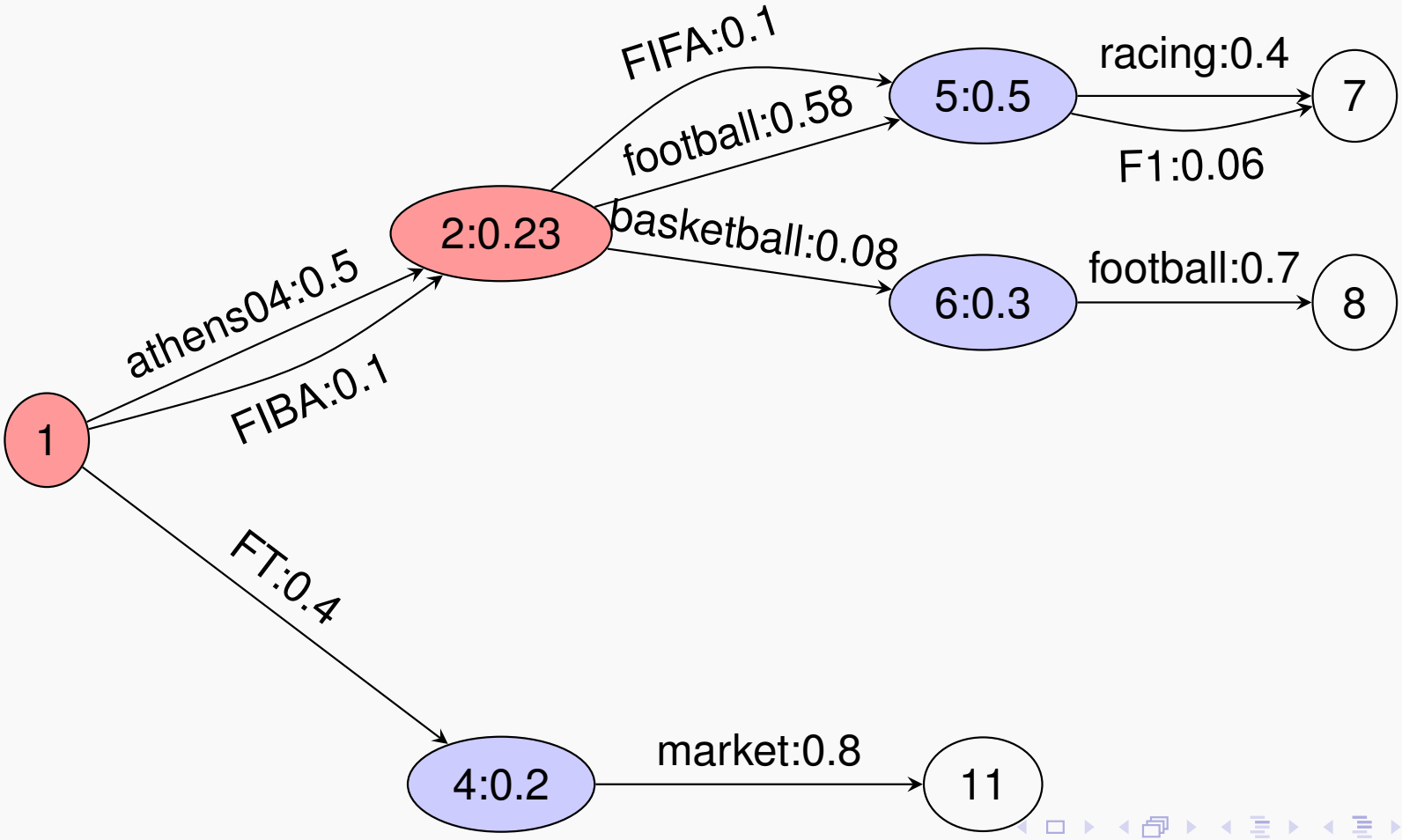
Content-Aware Navigation User Modeling with GI

A simple example



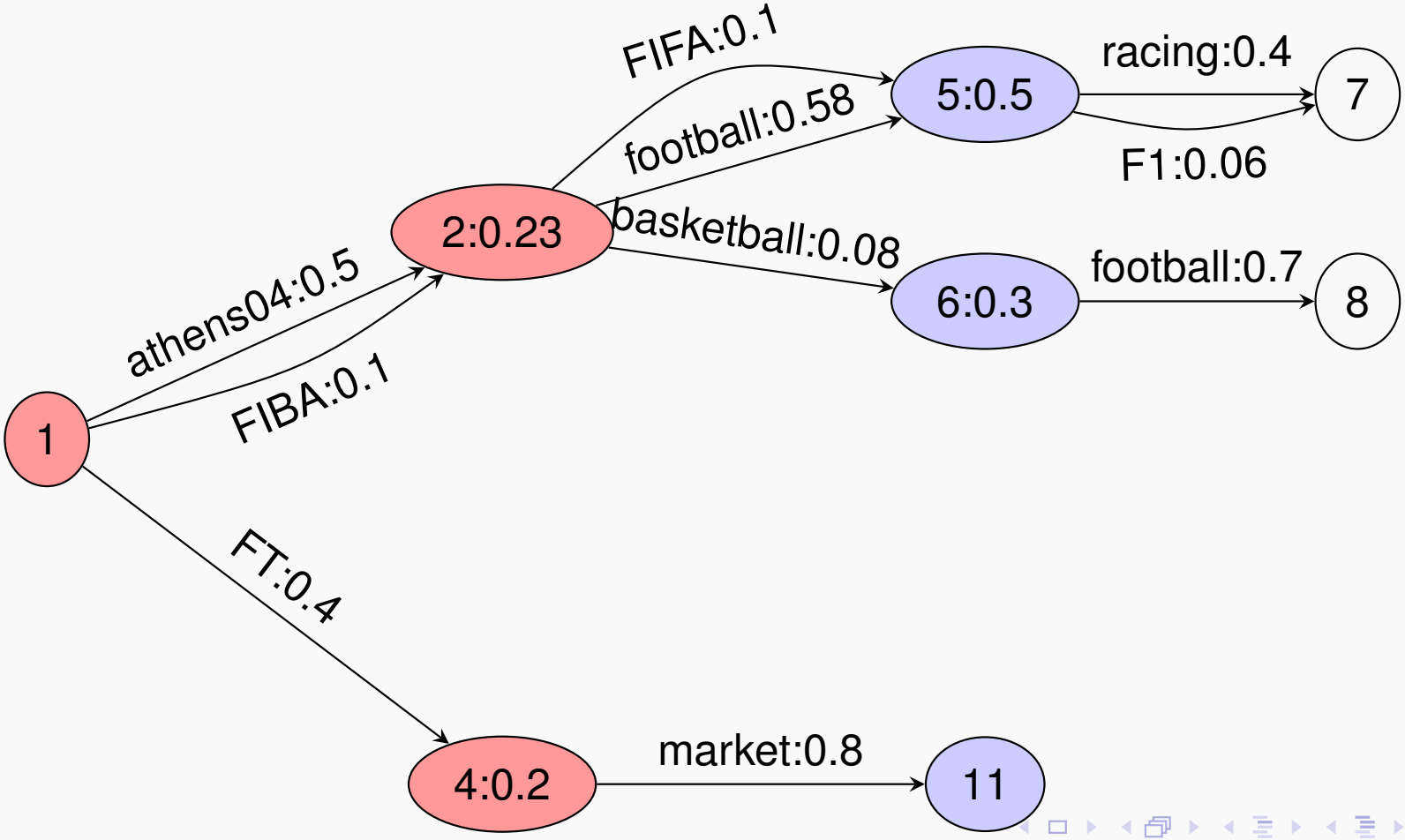
Content-Aware Navigation User Modeling with GI

A simple example



Content-Aware Navigation User Modeling with GI

A simple example



Content-Aware Navigation User Modeling with GI

On-line recommendation process

- Unlikely to trace a specific path of Web pages in the model.
- Modify recommendation process to use content similarity:
  1. Given a state s_i, with children S_i, and the next observed page of the user's session a, select $\arg\max_j sim(a, s_{ij})$.
  2. If $\arg\max_j sim(a, s_{ij}) < T_{sim}$ return to start state.
  3. At the end of the observed path, build recommendation list combining:
     - The transition probability to the final state's children.
     - The distance of each page in a state to the state's centroid.

Content-Aware Navigation User Modeling with GI

Evaluation

- Data: the ISP data used for personalized directories also.
- Modification of the Expected Utility measure:

$$EU_a = \sum_{j=0}^{n-1} \frac{sim(a, p_j)}{2^{j/h}}$$

- Comparison to content-only recommendation:
  1. Store all pages in the modeling phase.
  2. Score stored pages, according to average distance from the observed path.
  3. Produce a list of the n top-scoring pages.

Content-Aware Navigation User Modeling with GI

Results

method	EU
CANUMGI-A	8.57
CANUMGI-B	21.72
CANUMGI-C	20.59
CONTENT	24.25

Content-Aware Navigation User Modeling with GI

Results

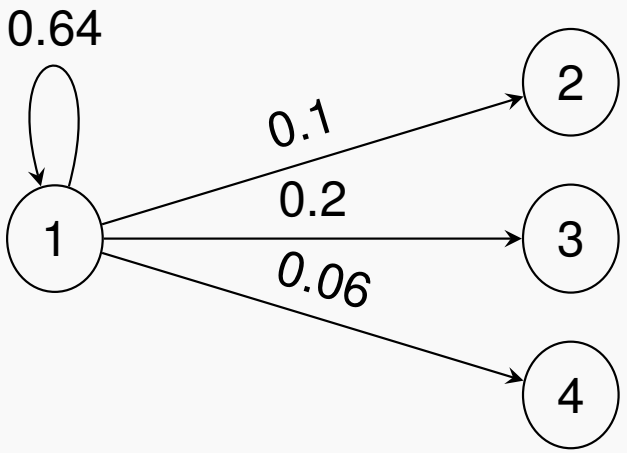
method	EU
CANUMGI-A	8.57
CANUMGI-B	21.72
CANUMGI-C	20.59
CONTENT	24.25

Does the navigation model help?

Content-Aware Navigation User Modeling with GI

Results

method	EU
CANUMGI-A	8.57
CANUMGI-B	21.72
CANUMGI-C	20.59
CONTENT	24.25

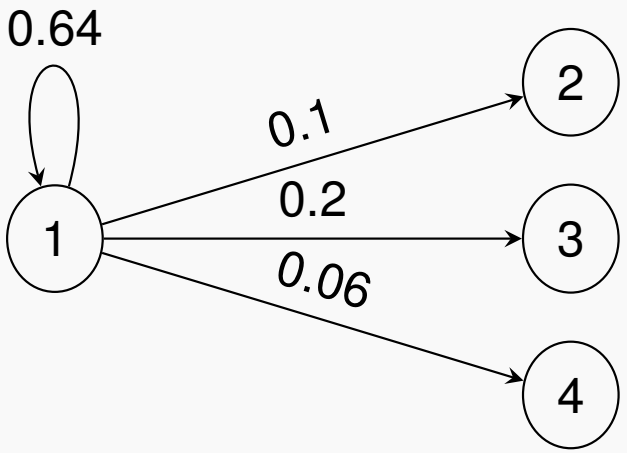


Does the navigation model help?

Content-Aware Navigation User Modeling with GI

Results

method	EU
CANUMGI-A	8.57
CANUMGI-B	21.72
CANUMGI-C	20.59
CONTENT	24.25



Does the navigation model help?

Navigation sequences are thematic.

Two facets of Web community discovery

- Discovery of Web user communities.
  - Analysis of usage data.
  - Discovery of interest and navigation patterns.
  - Communities of content consumers.
- Discovery of Web communities.
  - Analysis of Web structure.
  - Discovery of graph patterns (linkage of pages).
  - Communities of content creators.

Active Web users

- Web users are increasingly becoming content creators and service providers.
- At the same time they remain content consumers and service users.
- Active users are both creators and consumers.
- Many new services support active users:
  - Users as publishers, e.g. blogs, fora etc.
  - Collaborative creation of content and knowledge, e.g. flickr, del.icio.us, Yahoo!Answers, Wikipedia, bibsonomy, etc.

Community discovery

- Active user community discovery combines the existing approaches.
- Discovery needs to take into account:
  - Usage: what the user has chosen to see.
  - Content: what the user has contributed; how it relates to what the user read.
  - Structure: links between content created by different users.
- Active user community models combine this information into commonly observed patterns of community behavior.
- Discovery can also help evolve manually created communities.

Extending the Web

- Search engines:
  - Content creation and access on the mobile (e.g. Yahoo!Go).
  - Web as a medium of communication, even on the move.
- SensorPlanet (Nokia):
  - Mobile terminals as sensors providing user context.
  - Facilitate instant communities and nets, based on sensed locality and user profile.
- Ambient Semantics (MIT MediaLab):
  - Wearable RFID sensors to track things you pick up and people you meet.
  - Personal serendipity assistant:
    "What did my friends think of this book?"
    "What common interests do I share with this person?"
- and many others . . .

With or without the Web

- Digital switch-over in communication and broadcasting:
  - Traditional consumer services (e.g. TV) are becoming interactive.
  - New business opportunities for broadcasting and telecom providers to support active users.
- EU FP7 networked media:
  - Systems and application platforms to support media creation and management.
  - Support for individuals and self-organised creative communities.
- Applications remain mostly related to the Internet.

Community discovery

- Increased availability makes information itself less useful. It is just there.
- We need to answer new questions:
  - Where does the information come from?
  - Where and how should I contribute my content?
- Communities (local and global) become essential.
- Knowledge discovery can facilitate self-organising and dynamic communities.
- The KD approach is similar to the Web, but ...
- the nature and scale of the data is different.

Summary of our work so far

- Personalization is a major requirement for the Web.
- User modeling is a great challenge for Web personalization.
- Can we discover good models in usage data?
- We have developed methods for:
  - Discovering communities and stereotypes for a Web site.
  - Discovering navigation grammars for a Web site.
  - Personalizing Web directories.
  - Discovering navigation grammars for the Web.

Future directions

- Focus is on whole-Web personalization.
- Test further navigation-does-not-help hypothesis.
- Use Web directories to improve navigation modeling.
- Personalising Web search.
- Discovery of active user communities on the Web and beyond.

Tools, Systems and Applications

[PCHCI2001, KES2006, CROSSMARC, M-PIRO, INDIGO]

- KOINOTITES: a tool for community discovery
- PServer: A generic personalization server
- CROSSMARC: Personalized e-business (product comparison)
- M-PIRO, INDIGO: Personalized e-museum guidance
- PNS: Multi-source personalized news service

User Modeling 2007

11th International Conference on User Modeling (UM 2007)
Corfu , Greece , 25-29 June, 2007

Organized by
the National Center for Scientific Research "Demokritos",
in collaboration with
the Ionian University and User Modeling Inc.

http://www.iit.demokritos.gr/um2007/