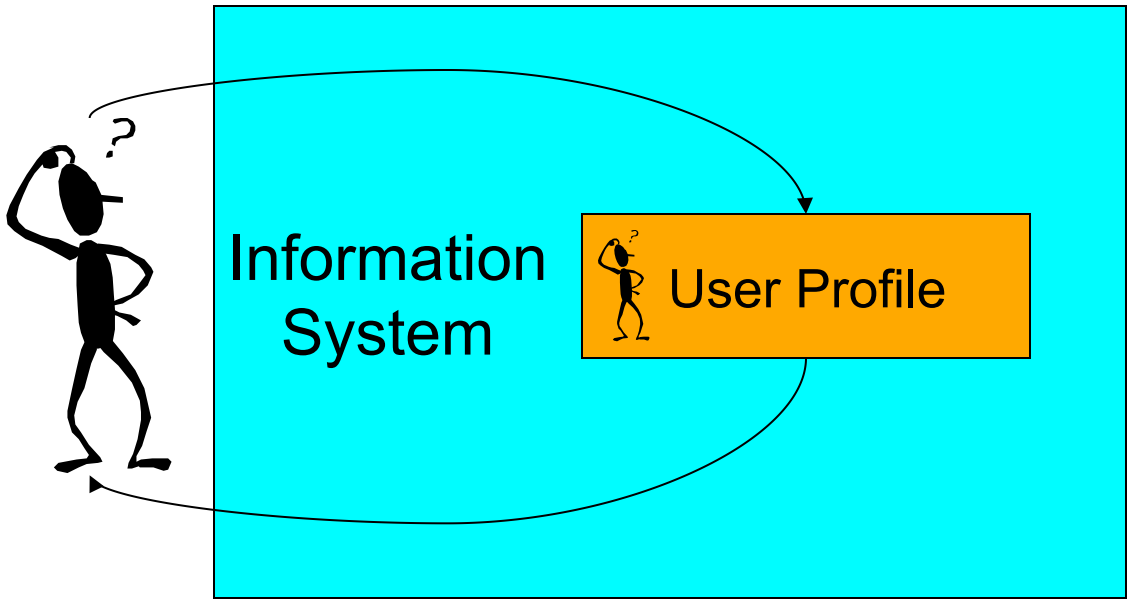
INFSCI 2480
User Profiles for Personalized Information Access

Peter Brusilovsky
School of Information Sciences
University of Pittsburgh, USA
http://www.sis.pitt.edu/~peterb/
With slides of Qiang Ye, INFSCI 3954 The Adaptive Web

Overview

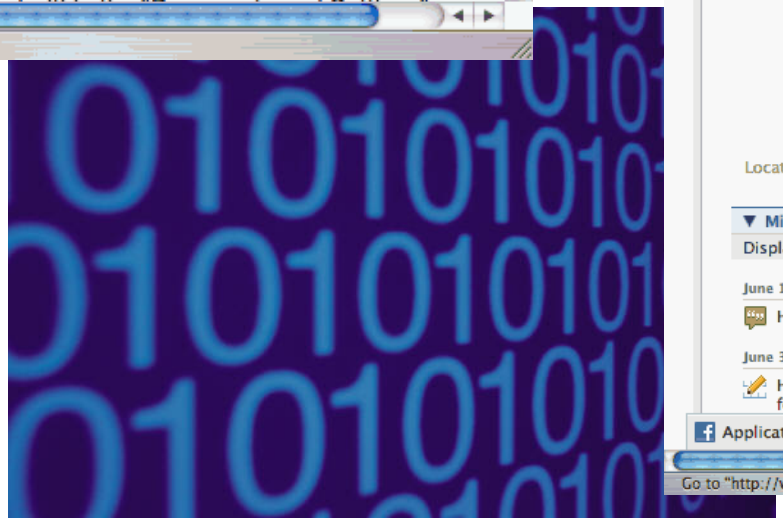
- Introduction
  - Definition
  - Classification
  - The Big Picture
- Information Collection
- User Profile Representation
- User Profile Construction
- Issues

User Profiles



User profile is a representation of a user in an information system

What is User Profile?



User Profile

- Common term for user models in information retrieval, filtering, and content-based recommender system
- A user's profile is a collection of information about the user of the system, which the system collects and maintains in order to improve the quality of information access
- User profile is applied to get the user to more relevant information

SDI: The Origin of Profiles

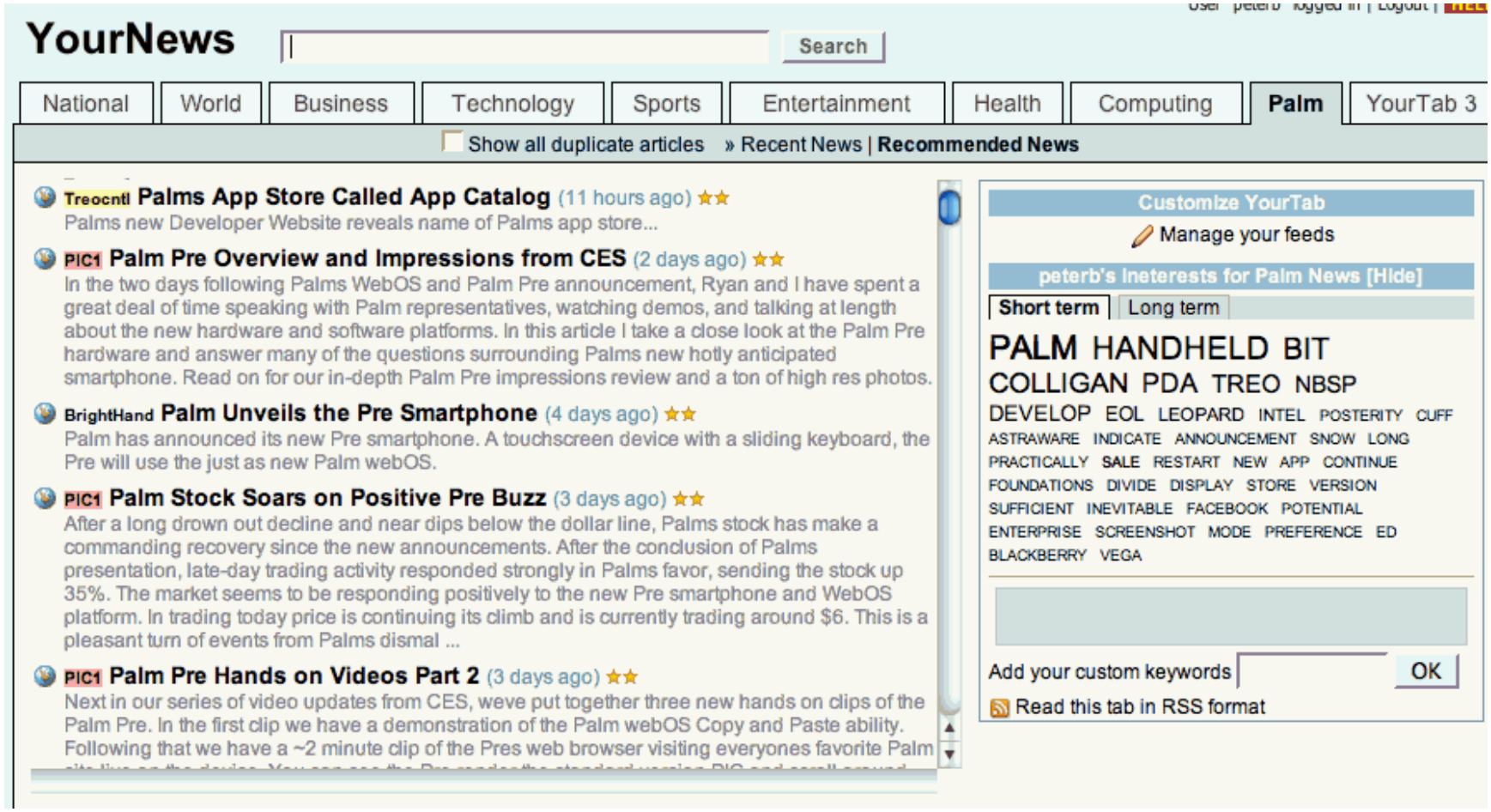
- Selective Dissemination of Information (SDI)
  - User defines her profile of interests
  - System filters all relevant new sources

    "Artificial intelligence" and "education"
- Profile - while looks like a query - is really *more* than a query since it represents long term interests
  - that is where the work on user profiling started
- Used for retrospective search and awareness
- Profiles kept updated by the users

Core vs. Extended User Profile

- Core profile
  - contains information related to the user search goals and interests
- Extended profile
  - contains information related to the user as a person
    - demographic information, e.g., name, age, country
    - education level
    - abilities
    - profession…
- Determined by the application needs

Example: Core User Profile in YourNews

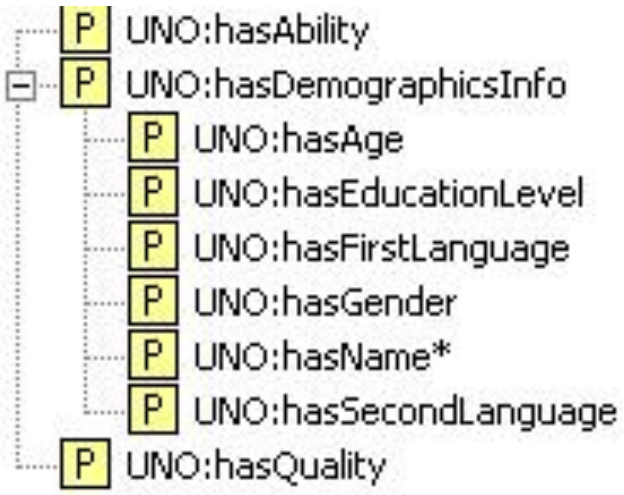


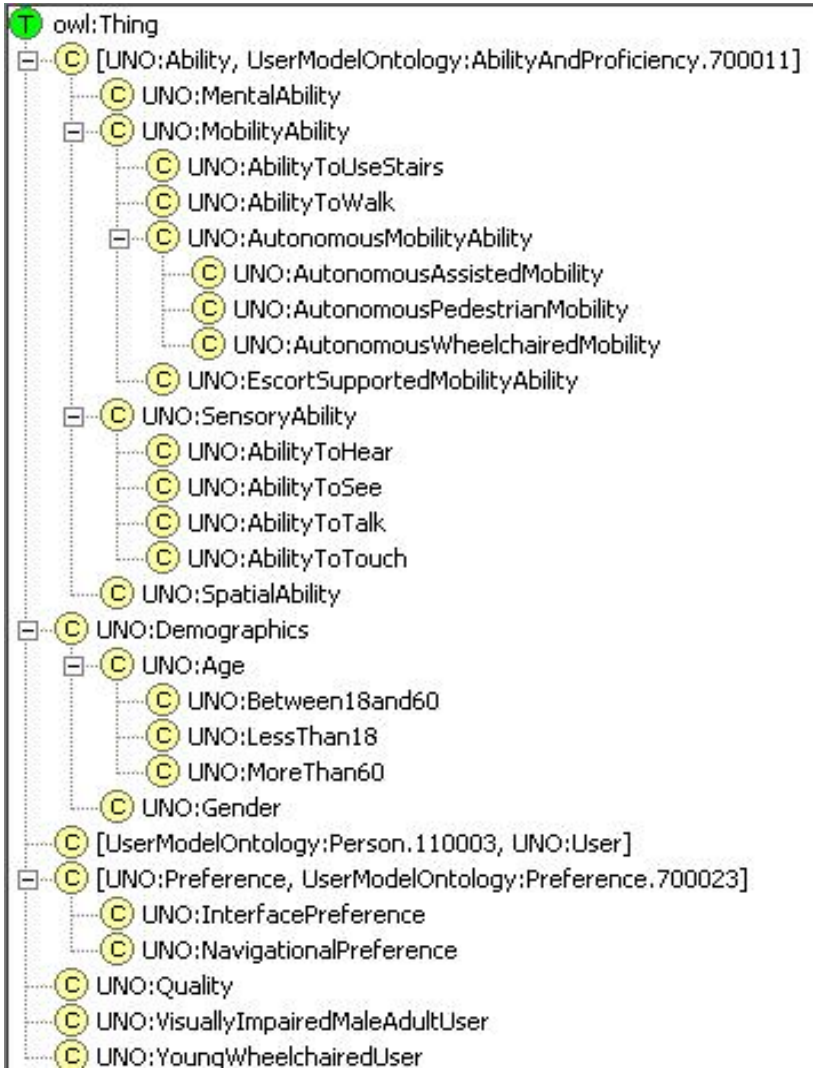
http://amber.exp.sis.pitt.edu/yournews/

Example: Extended User Profile in a Navigation Systems UNO

Classes

Properties

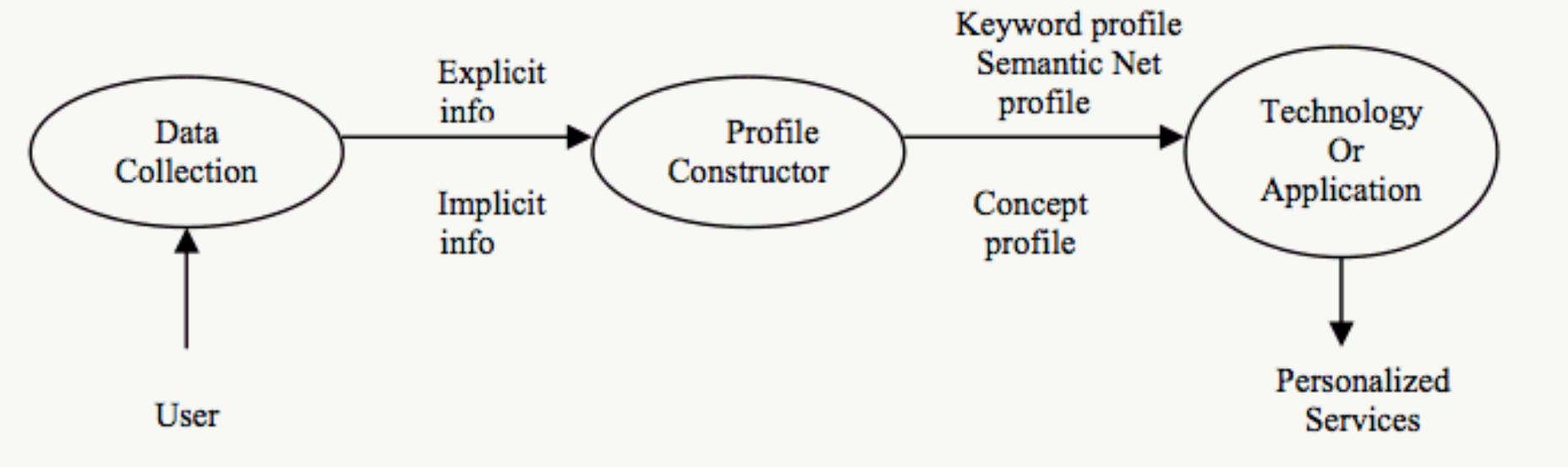




User Profiles: Classification

- According to the way information is collected:
  - explicit, through user intervention
  - implicit, through agents that passively monitor user activities
- According to the life-period of the profile
  - Static profiles that maintain the same information over time.
  - Dynamic profiles that can be modified or augmented.
    - Short-term profiles represent the user's current interests
    - Long-term profiles indicate interests that are not subject to frequent changes over time
- Structure
  - Keyword profiles
  - Semantic net profiles
  - Concept profiles

The Big Picture



Overview

- Introduction
- Information Collection
  - User Identification Method
  - User Information Collection Method
- User Profile Representation
- User Profile Construction
- Privacy Issue

User Identification

- Five basic approaches to user identification:
  - software agents
  - logins
  - enhanced proxy servers
  - cookies
  - session ids
- The first 3 techniques are more accurate, but require active participation of the user. The last 2 are less invasive

User Identification - Intrusive

- Software agent
  - a small program residing on the user's computer, collecting their information and sharing this with a server via some protocol.
  - Pros: the most reliable because of full control over the implementation of the application and the protocol used for identification.
  - Cons: it requires user-participation in order to install the desktop software. And if the user uses a different computer, no user information will be collected.
- Logins
  - Pros: Accurate, reliable, can use the same profile from a variety of physical locations with different computers.
  - Cons: user must create an account via a registration process, and login and logout each time they visit the site
- Enhanced proxy servers
  - Pros: provide reasonably accurate user identification.
  - Cons: require that the user register their computer with a proxy server. Thus, they are generally able to identify users connecting from only one location.

User Identification - Nonintrusive

- Cookies
  - The first time that a particular IP address connects to the system, a new user id is created, and stored in a cookie on the user's computer. When they revisit the same site from the same computer, the same user id is used.
  - Pros: no burden on the user at all.
  - Cons: if the user uses more than one computer, each computer will have a separate cookie, and thus a separate user profile. Also, if the computer is used by more than one user, and all users share the same local user id, they will all share the same, inaccurate profile. Finally if the user clears their cookies, they will lose their profile altogether.
- Session IDs
  - Similar to cookies, but there is no storage of the user-id between visits . Each user begins each session with a blank profile, but their activity during the visit is tracked.
  - Cons: no permanent user profile can be built, but adaptation is possible during the session.

User Information Collection

- Explicit Feedback Systems
  - Rely on direct user intervention, typically via HTML forms.
  - More accurate, but place extra burden on users
  - The data collected may contain demographic information such as birthday, marriage status, job, or personal interests.
  - Users may not choose to participate or accurately report their interests. Profiles remain static while user interests may change over time
- Implicit Feedback Systems
  - Collect user information while user is performing regular tasks
  - For open Web personalization require additional software to capture user activity

What Kind of Implicit Feedback?

- Better tracking of regular (reading) activities
  - Time spent
  - Scrolling and mouse movement
  - Eye tracking
- Enabling and tracking additional interest-bearing activities
  - Bookmarking
  - Downloading (Pazzani's paper recommender)
  - Annotating (Knowledge Sea)

IF Collection: Browser Cache and Proxy Server

- Browsing histories can be collected in two ways
  - users share their browsing caches on a periodic basis
  - users install a proxy server that acts as their gateway to the Internet, thereby capturing all Internet traffic generated by the user (iSpy operates this way)
- Disadvantages:
  - Sharing histories requires too much work from the user
  - Browsing histories are typically shared with one particular Web site, allowing that site only to provide personalized services.
  - Typically collects browsing history from a *single computer*. What if user uses multiple computer?
    - Share browsing cache from multiple computers
    - Install same proxy server on each computer
    - Use a login system with same user profile

IF Collection: Browser Agent

- Implemented as either a standalone application that includes browsing capabilities or a plug-in to an existing browser (i.e., Alexa, HeyStaks)
- Advantage:
  - Collects richer information about the user. In addition to browsing history, the agents can also collect actions performed on the Web page such as bookmarking, downloading, scrolling and mousing.
- Disadvantages:
  - Requires users to install a new application or plugin on their computers
  - Requires a large investment in software development and maintenance
  - Since it is resident on a personal computer, the user profile built would typically only be available when the user was using that particular computer
    - Or install on multiple computers and assure synchronization (HeyStaks)

IF Collection: Desktop Agents

- The searches is not limited to the Web, but they would also include databases to which the user has access, and the users personal documents. Such search systems are implemented in tools like Google Desktop Search.
- The information found in the personal documents and databases could be used to enhance the user profile
  - Server-side approach collect only the activities the user performs while interacting with the site providing the personalized services.
- Desktop agents are essentially client-side approaches and may place some burden on the users in order to collect and/or share the log of their activities unless tightly integrated with OS
  - Microsoft, Apple, and Google are actively working on it

IF Collection: Web and Search Logs

- Web logs capture the browsing histories for individual users at a given website
  - Can be used to adapt website organization based on user behavior.
- Search logs contain info about queries from a particular user and date/time/result of the query
  - Can be used to build user profiling to help personalized and social search
- Advantage: user does not need to install a desktop application and/or upload their information to the personalized service.
- Disadvantage: only the activities at the search site itself are tracked, much less information is available
- Heavily used by IBM, Google and Microsoft

Overview

- Introduction
- Information Collection
- User Profile Representation
  - Keyword profiles
  - Semantic Network Profiles
  - Concept Profiles
- User Profile Construction
- Issues

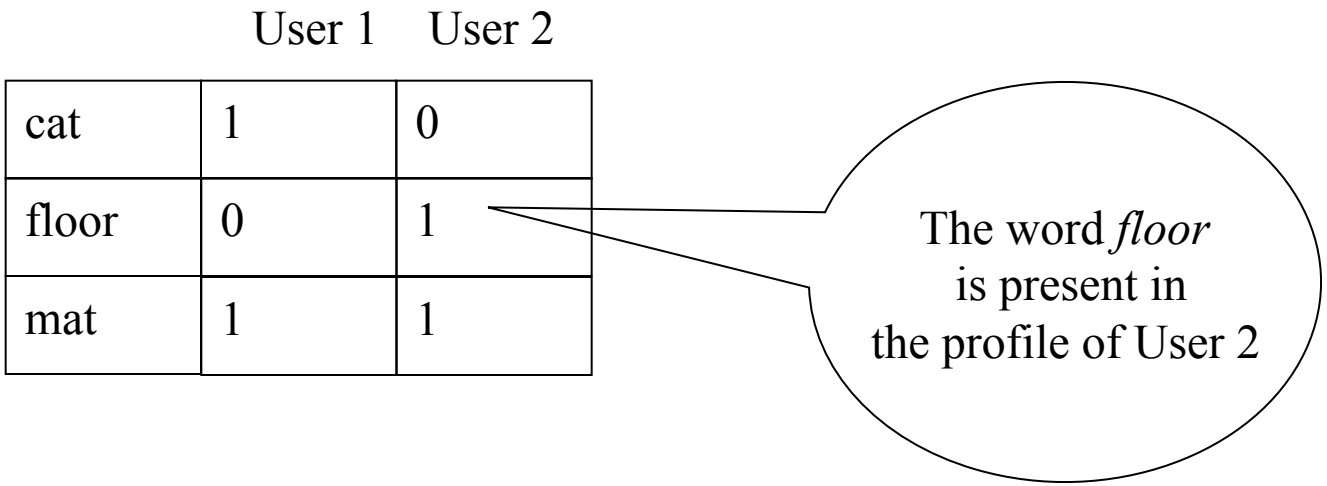
Keyword Profiles

- Based on keywords extracted from web pages visited, bookmarked, saved or explicitly provided by the user
- Bag-of-words
  - Simply a set of most popular words, can be used in different kinds of systems
  - Each keyword may be also associated with a numerical weight representing its importance in the profile
- Profile vectors
  - An overlay of a keyword vector used in document modeling in a specific system
  - 0-1 vector
  - Weighted vector
- Benefits
  - Simplicity
- Shortcomings:
  - Words may have multiple meanings. Same idea can be expressed by different words. Because of this polysemy and synonymy, the keywords in the user profile are ambiguous, making the profile inaccurate

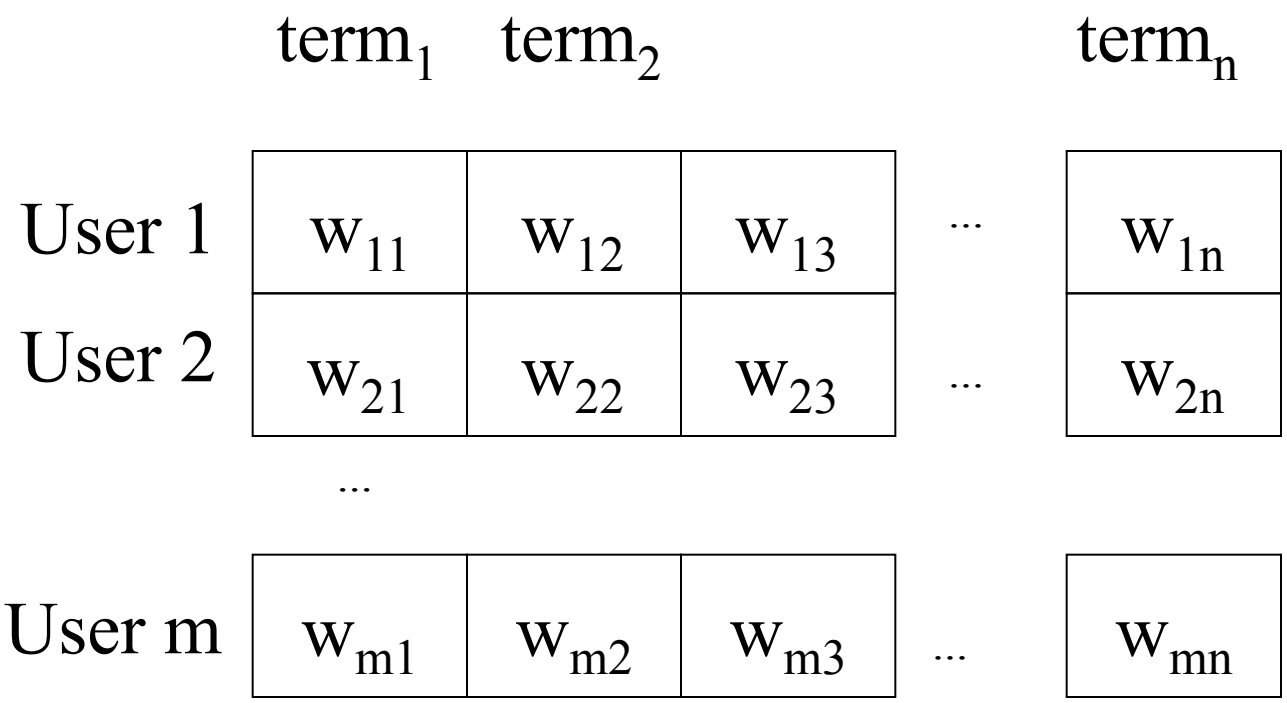
0-1 Keyword profile

- Rows represent document terms
- Columns represent users

`User 1 liked document “the cat is on the mat”`
`User 2 liked document “the mat is on the floor”`



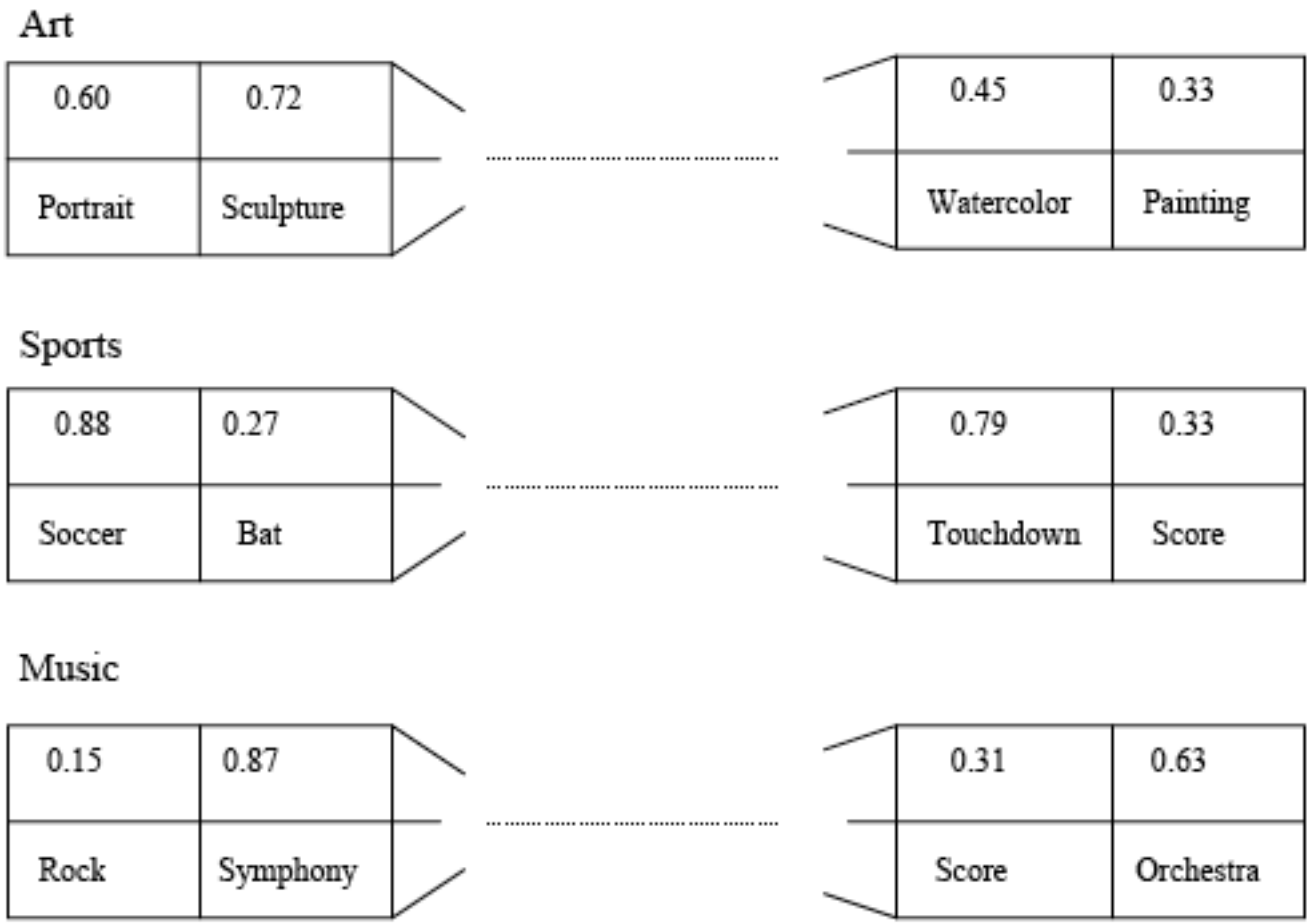
Weighed Keyword Profile



Advanced Keyword Profiles

- Dealing with shortcomings: synonymy, polysemy, interest drift
  - In *PEA* project, rather than creating a single profile for the user, the user is represented as a set of keyword vectors, one per bookmark (interest)
  - *Alipes* expands this approach by representing each interest with three keyword vectors, i.e., a long-term descriptor and two short-term descriptors, one positive and one negative
- These approaches are complementary
  - *YourNews* keeps separate profiles for each topic (tab) *and* distinguish short and long-term profiles

Domain-Based User Profile



Semantic Network Profile

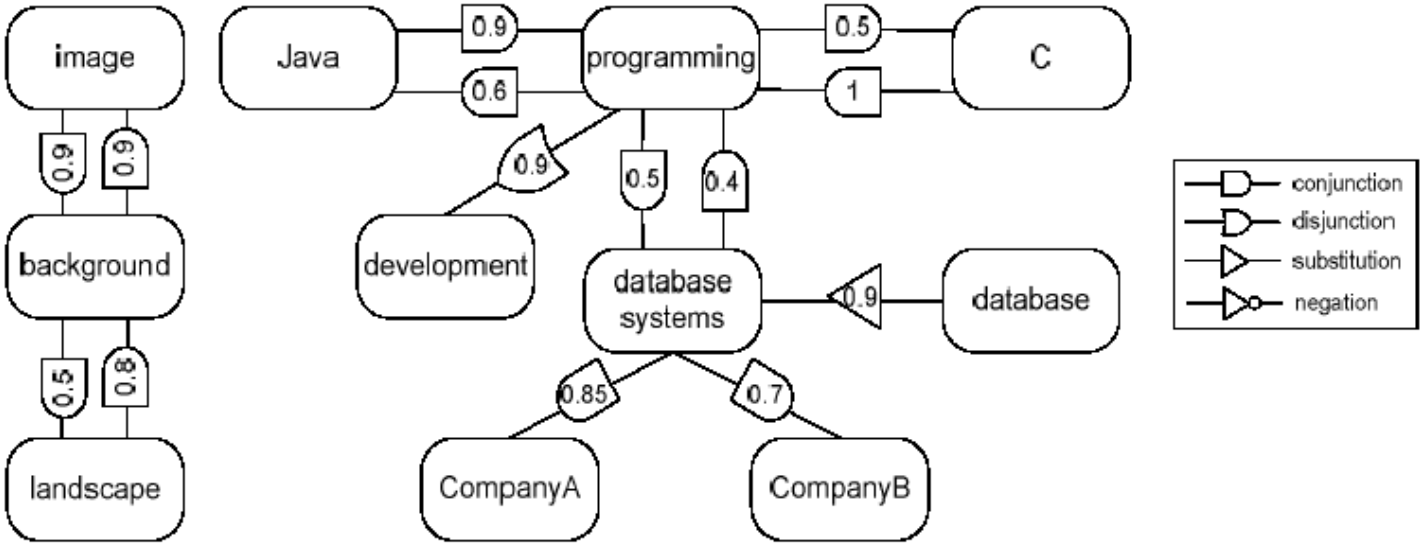
- To address the polysemy problem in keyword-based profiles, the profiles may be represented by a weighted semantic network in which each node contains a particular word found in the corpus and arcs are created representing co-occurrences of the two words in the connected nodes.

- In SiteIF project, they found that representing individual words as nodes in semantic network is not accurate enough to discriminate word meanings. Instead, they group related words together in “synsets”.

- A user profile is a semantic network where the nodes are “synsets”, the arcs are co-occurrences of the “synsets” members within a document of interest to the user, and the node and arc weights represent the users level of interest.

Semantic Network Profile

- Advanced relevance network for query expansion
- `java -> java and programming -> java and (programming or development)`



A Unified User-Profile Framework for Query Disambiguation and Personalization Georgia Koutrika and Yannis Ioannidis, `http://adiret.cs.uni-magdeburg.de/pia2005/Proceedings.htm`

Concept Profile

- Similar to semantic network-based profile with nodes and arcs. But the nodes represent abstract topics considered interesting to the user, rather than specific words or sets of related words.
- It is suggested using hierarchical concepts, rather than a flat set of concepts, to enables generalizations. The simplest concept hierarchy based profiles are constructed from a reference taxonomy (WordNet) or thesaurus. More complex profiles may be constructed from reference ontology (ODP).
- The levels in the concept hierarchy can be fixed, or they can change dynamically according to the users interests.

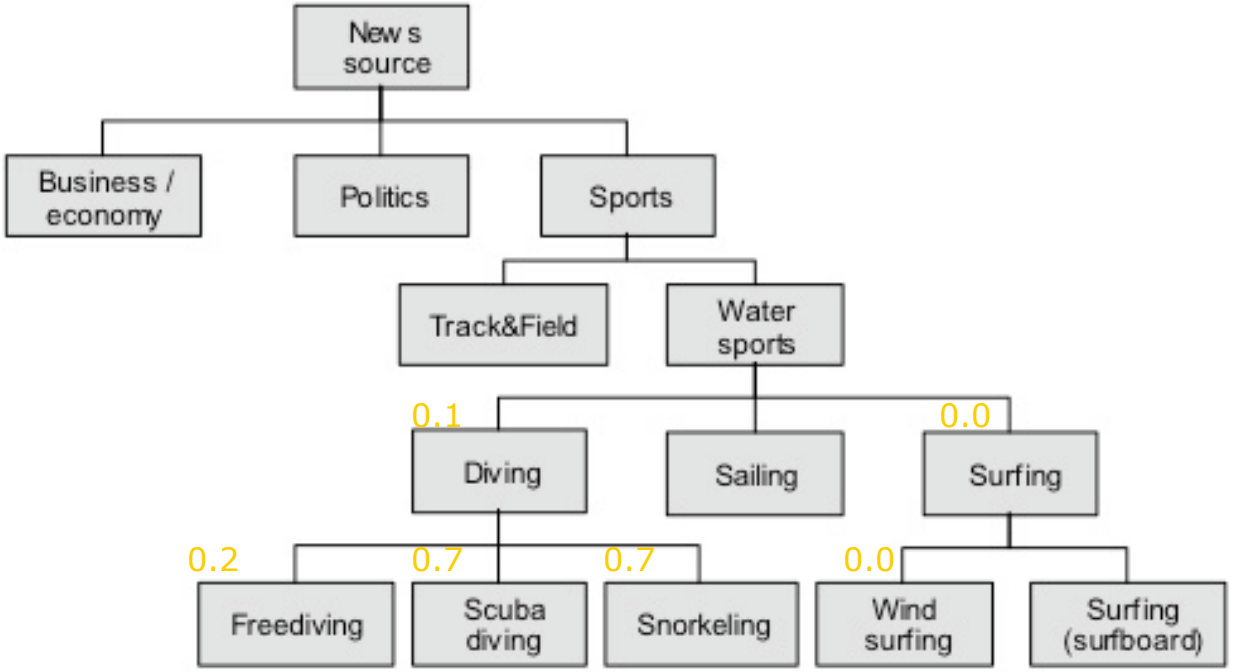
Concept Profiles

- Because creating a broad and deep concept hierarchy is an expensive, mostly manual process, profiles are typically based on subsets of existing concept hierarchies.

- When using an existing directory as a source of concepts, certain transformations must take place to turn directory contents into a concept hierarchy.
  - Usually only top 3 levels are used.
  - Discard those subjects with too few associated Web pages to act as examples for training

Concept profile over news taxonomy

- For each domain concept or taxon an overlay model stores estimated level of interests



Overview

- Introduction
- Information Collection
  - User Identification Method
  - User Information Collection Method
- User Profile Representation
- User Profile Construction
  - Building Keyword Profiles
  - Building Semantic Network Profiles
  - Building Concept Profiles
- Issues

Building Keyword Profiles

- Keyword-based profiles are initially created by extracting keywords from Web pages collected.
- keyword weighting is done to identify the most important keywords from a given Web page. Most popular weighting scheme: tf*idf from information retrieval theory.
- In addition to the tf*idf, other projects have explored using Latent Semantic Indexing (LSI) and Linear Least Squares Fit (LLSF) for creating the keyword-based feature vectors.
- The number of words extracted from a single page is frequently capped: only the top *N* most highly weighted terms from any page contribute to the profile.

Building Keyword Profiles - example

- *Alipes project* creates user profiles that are based upon interests. Each interest is modeled by three keyword vectors: long-term; short-term (postitive), and short-term (negative).

- The creation of new interests is based on a similarity threshold. When a document vector is added to the user profile, it is compared to each of the three vectors for each interest using the cosine similarity metric.
  - If the similarity exceeds a threshold, the document vector is added to the best matching interest.
  - If, there is no sufficient match, a new interest is created and seeded with the document vector

Building Semantic Network Profile

- The keywords are added to a semantic network
  - If the keyword is already in the semantic network, that node's score is increased by the value of the user's feedback (or decreased, if the feedback is negative).
  - If the keyword does not already appear, then a new node is created.
  - Finally, the set of keywords are used to update the weights on the co-occurrence arcs.

Building Semantic Network Profile

SiteIF project

- Learns user's interests from implicit feedback.
- Keywords are extracted from web pages, and mapped into synsets using WordNet. Polysemous words are then disambiguated by analyzing their synsets to identify the most likely sense given the other words in the document.
- Finally, the synsets are combined to yield a user profile that is a semantic net whose nodes are synsets and arcs between nodes are the co-occurrence relation of two synsets;
- every node and every arc has a weight. The weights of the net are periodically updated. Nodes and arcs that are no longer useful may be removed from the net.

Building Concept Profiles

Persona project

- Initially, user profiles are represented as a collection of weighted concepts based on the Open Directory Project's concept hierarchy.
- As the user searches the collection of pre-classified documents in the ODP, they are asked to provide explicit feedback on the resulting pages. This feedback is then used to update their profile.
- Because Persona uses pre-classified documents, the profile is able to contain any concepts in the ODP and the mapping of visited pages to concepts is very accurate.

Building Concept Profiles

Obiwan Project

- Represents user profiles as a weighted concept hierarchy built from a reference ontology (ODP).
- But it is not restricted to building the user profiles from pre-classified documents. Any source of representative text may be automatically classified by the system to find the best matching concepts from the ODP, and then those concepts have their weights increased.
- Using text classification to map the user information into the appropriate concept in the hierarchy. Several different text classification methods have been used for comparing the new documents to the reference set, such as SVM, KNN, Naïve Bayesian, Decision Tree and Neural Networks.

Overview

- Introduction
- Information Collection
  - User Identification Method
  - User Information Collection Method
- User Profile Representation
- User Profile Construction
- Issues
  - Privacy
  - Profile exchange
  - Profile editing

Privacy Issues

- Personal user information is critical data and careful attention should be given to where and how user profiles are stored.
  - User might prefer to store their information on their local machine or they may not want their personal information stored at all.
- All personal information must be protected and, users should be allowed to view and modify their personal information.
- User's real identity is not necessary, many countries protect the privacy of identified or identifiable users.
- User identification can be obtained using mechanisms such as session ids or cookies that provide anonymity. Even methods requiring a login process can be anonymous if users are be allowed to use pseudonyms rather than their true identity.

Profile Exchange

- Multiple systems collect user profiles
- Integrating and exchanging profiles could lead to better personalization
- New stream of research on Ubiquitous User Modeling
- Ontologies for profile exchange
  - GUMO
  - UNO

Who Maintains the Profile?

- Profile is provided and maintained by the user/administrator
  - Sometimes the only choice
- The system constructs and updates the profile (automatic personalization)
- Collaborative - user and system
  - User creates, system maintains
  - User can influence and edit
  - Does it help or not?

Conclusions

- An accurate representation of a user's interests is crucial to the performance of (content-based) personalized information access systems
- We surveyed some of the most popular techniques for collecting user information, representing and building user profiles
- On-going research topics…
  - How to improve profile accuracy?
  - How to quickly achieve profile stability? How to identify major/minor, long-term/short-term interest of users? How to determine appropriate level of depth in the interest hierarchy in user profile…?