Clustering for Opportunistic Communication

Jay Budzik

Intelligent Information Laboratory
Department of Computer Science
Northwestern University
Collaborators

- Shannon Bradshaw
- Xiaobin Fu
- Kris Hammond
Broad Context

- Building software that can proactively help you achieve your goals by understanding enough about what you’re doing
- Focus: facilitating resource awareness
  - Watson (documents)
  - I2I (potential collaborators)
Watson

- Watson allows you to easily maintain an awareness of relevant online documents in the context of your work
- See paper, IUI 2000
In both Citizen Kane and The Magnificent Ambersons Orson Welles


to do what is impossible for many theater directors who move into film-making, managed simultaneously to incorporate the literary nature of the theater and visual effects available to him in the medium of film. The storytelling craft of theater is evident in both these films. What is interesting to the theater student might be termed an exploitation of theatrical form. His knowledge of Shakespeare and the Ancient Greeks seems to have greatly informed his craft, sometimes for effect, and sometimes serving as a detriment.

Turning first to Citizen Kane, one first notices that Welles has played an


epic form. He tells the story of a man born to greatness, his great deeds, and due to what might be the hamartia - tragic flaw. Although tragic flaw might be a reductive or simplistic term, in this case, like Shakespeare before him, taken what he needed from the Greeks and moved it into his own particular...
Intelligent Information Laboratory
at Northwestern University

Summary

Library
A trip to the movies: 100 years of film as art. (includes related article on films released during the 1960s)(Cover Story)
The Humanist (Hinnichs, Bruce) 03-11-1966; 7(7)

Pictures
Image from Ditto.com
www.homevideos.com/...12.htm (Ditto)

News
CNN - 'The Film 100' a history lesson for film buffs - August 5, 1999
ATLANTA (CNN) -- Let's get this out in the open. Most people won't agree with the following list. Not at first, anyway. It's designed to spark debate and interest -- and...

Search in context: 
Status: Processing results...
Watson --> I2I

- Watson is all about tracking and using context to drive proactive media retrieval
- I2I is aimed at fostering informal collaboration and communication through awareness of shared contexts
The Basic Idea

- I2I connects users based on the work they are performing
- For example, I2I users writing papers on a similar topic can
  - become aware of each other’s activities through the system and
  - use this awareness as a starting point for collaboration
- We want to make traditionally solitary activities more collaborative by embedding context-sensitive activity awareness facilities into everyday applications
Many of our conversations are not planned in advance

- Awareness of others who are situated in a similar context facilitates informal collaboration and communication
  - E.g., BOF sessions, lunch at conferences

- CMCs typically leave context-awareness out
  - You have to know “where to go” and decide to go “there” to find people to talk with
  - Fixed, place-based metaphors

- And so the cost of finding help/collaborators often outweighs the perceived benefit
Opportunistic Communication

- We call communication that arises out of an awareness of shared context *opportunistic*
  - Awareness of common active goals (or immediate interests) is required for people to help each other
  - We want to promote this kind of awareness by tracking the work people do and noticing opportunities for collaboration
Clustering Work Contexts

- Similarity-based clustering offers a mechanism for discovering common work contexts
  - Work contexts can be represented as feature vectors
  - Neighborhoods of similar work contexts --> communities of common interest

- What is the content of the feature vector?
  - Goals/Plans
  - Process representations
  - Textual representations
Documents are a Window into a User’s Goals and Interests

- People manipulating similar documents
  - On the Web
  - In a word processor
  - In any document manipulation application, in general
- The idea is that people manipulating similar documents often have common goals
  - Obviously this is not perfect: telephone book vs. focused report
  - But the system is opt-in
Richard M. Nixon

Thirty-Seventh President

Fun Fact: Though President Richard Nixon disliked making the White House swimming pool filled in, to give reporters more room for events. But one president’s decisions about the White House pool often lead to the next. After Richard Nixon resigned on August 9, 1974, the next president, his friends had another pool dug on the White House grounds.

Fast Fact: The Watergate scandal forced Richard M. Nix...
Richard M. Nixon

Thirty-Seventh President

[Patrick]

Fun Fact: Though President Richard Nixon disliked the White House swimming pool filled in, to give reporters more news events. But one president's decisions about the White House, the next. After Richard Nixon resigned on August 9, 1974, the president, his friends had another pool dug on the White

Fast Fact: The Watergate scandal forced Richard M.
Richard Nixon

Thirty-Seventh President

Fun Fact: Though President Richard Nixon disliked modernizing the White House swimming pool filled in, to give reporters more room for new events. But one president’s decisions about the White House’s pool size can change next. After Richard Nixon resigned on August 9, 1974, as the first president, his friends had another pool dug on the White House grounds.

Richard M. Nixon

Thirty-Seventh President

Fun Fact: Though President Richard Nixon disliked媒介 House swimming pool filled in, to give reporters more media events. But one president's decisions about the White House next. After Richard Nixon resigned on August 9, 1974, the president, his friends had another pool dug on the White

Richard M. Nixon

Thirty-Seventh President

Fun Fact: Though President Richard Nixon disliked maintaining the White House swimming pool filled in, to give reporters more access to the events. But one president’s decisions about the White House pool cost him his election. After Richard Nixon resigned on August 9, 1974, his successor, Gerald Ford, ordered the pool filled in.

Fast Fact: The Watergate scandal forced Richard M. Nixon to resign as president.
what do you think was on the missing nine minutes?

I don't know ...

His election in 1968 had climaxed a career unusual on two counts: his early success and...
Richard M. Nixon

Thirty-Seventh President

Fun Fact: Though President Richard Nixon disliked men's bathrooms, he had the White House swimming pool filled in, to give reporters more room to move around for press conferences. But one president's decisions about the White House bathrooms were soon reversed, and the House of Representatives voted to impeach him for his role in the Watergate scandal. He was the first president to resign from office.

Fast Fact: The Watergate scandal forced Richard M. Nixon to resign before his term ended. He was the third president to be impeached.
Richard M. Nixon

Thirty-Seventh President

Fun Fact: Though President Richard Nixon disliked most reporters, he agreed to let the White House swimming pool filled in, to give reporters more room and privacy. But one president's decisions about the White House pool can be quite subjective. If Richard Nixon had lived a little longer, he might have found out that his pool was dug to accommodate the next. After Richard Nixon resigned on August 9, 1974, he lived as a private citizen in San Clemente, California, until he died in 1994. In 1995, Reagan became president, his friends had another pool dug on the White House grounds.

Fast Fact: The Watergate scandal forced Richard M. Nixon to resign as President of the United States.
Contexts of use

- Community building across cultural/physical boundaries
  - In education
- Reduce replication, aid in expertise location, facilitate synchronization
  - In business, especially for large organizations
I2I Architecture

Clustering for Opportunistic Communication: Budzik, Bradshaw, Fu and Hammond
Northwestern University Computer Science
Brokering Opportunities for Collaboration

- A central broker computes a similarity matrix for user contexts.
- By grouping conceptually similar contexts together, I2I makes it more likely that people will see each other.
Context (Document) Similarity

- Vector-space model (Porter stemmer)
- TFIDF term weights
- Cosine measure (all due to Salton, et al.)
- Fixed similarity threshold
- Basically, if two documents have enough content-bearing words in common, they are deemed ‘similar enough’
Secondary objects are associated with Contexts

- People
- Chat Rooms
- Calling cards
- In the future, experts, representations of expertise (FAQs), and open questions
Keeping track of appropriate contexts for presentation

- Indexing calling cards and chat rooms in this way allows the system to maintain consistent relationships between the contributed content and the contexts in which a note was intended to be viewed.
- It accounts for the ephemeral nature of information on the Web (sites go down, pages move, and content changes), allowing the system to maintain correspondence between context and contributed content.
  - Similar to Bob Wilensky’s “robust hyperlinks” work (WWW9)

Clustering for Opportunistic Communication: Budzik, Bradshaw, Fu and Hammond
Northwestern University Computer Science
Prior Work

- Introduce visitors on the *same Web* page.
- Sociable Web (Donath, WWW2) and others
- Gooey, Odigo and others
- Why might document clustering be better?
  - Clustering makes connections more likely
  - How many people do we need to start to see results? (Grudin’s “critical mass” problem)
    - Is there a similarity threshold that optimizes the tradeoff between finding someone and finding someone *relevant*?
The Data

- Two days of browsing logs from 11 people in and outside of the department
  - Internet Explorer plug-in recorded content (not just URL) to a data file when a page was fully loaded
- 1612 unique URLs accessed 5039 times over 2 days
Access Frequencies vs. Frequency Rank
The Data

- Follows Zipf distribution, and mirrors characteristics of a larger data set (575K URLs, 591 users over ~3 months):
  - Cunha, Bestavros, and Corvella (BU CS TR-95-010)
- Large number of pages are accessed infrequently
- This implies there will be critical mass problems for page-based systems
  - All or nothing
  - Clustering based on user contexts may provide a solution
Simulating Large Numbers of Users

- Simulated users created by randomly sampling from the original distribution and averaging over 100 samplings
  - E.g., it was more likely that one of our simulated users would be at the more popular pages
  - We will do studies with larger numbers of users when the system is deployed
- Number of simulated people vs. average number of people they would see from a particular page

Clustering for Opportunistic Communication: Budzik, Bradshaw, Fu and Hammond
Northwestern University Computer Science
Number of people that must be using the system to see one other person on average, vs. threshold (interception of the regression line with $y = 1$).

29% improvement over URL-based systems
Effectiveness

- Clustering effectiveness all over the literature
- For this task
  - For each threshold (0.1 to 0.9)
    - 10 random ‘source’ documents
      - 10 random ‘similar’ documents
- 900 comparisons by a single volunteer
  - Forthcoming study (HT02) on variance of similarity judgments suggest single-subject design is still representative
Threshold vs. percentage of inappropriate associations made by the system. As expected, as the threshold increases, the number of erroneous associations decreases.
Balancing the Tradeoff

- Fixed similarity threshold of about 0.7
  - Adaptive similarity
  - New representations (e.g., global and local history) for adaptive computation
    - Use more detailed representations to discriminate among visitors at more popular sites
  - Clustering improves the chances a given user will see some relevant other
Clustering User Contexts as a Basis For Awareness

- Provides a framework in which constraints can easily be manipulated by the system so a manageable number of people can be presented
- Provides a framework for including novel representations of user contexts
  - Document Contents + Historical Profile
  - Other domains (e.g., CAD, music, etc.)
Textual Representations of Context

- Much of work is document-centric
- Unlike URL-based representations, text allows unpublished documents to serve as an entry point to the system
  - People who are writing can see others who are viewing related items on the Web
- Account for the multiplicity of documents on the same topic and access patterns observed on the Web
- Doesn’t help for the most popular pages
Conclusion

- I2I embeds communication facilities in applications so that users that share interests can be aware of each other and communicate freely in an informal environment.
- I2I proposes a framework for opportunistic communication that overcomes many of the problems associated with document-based awareness and annotation technologies.
- New method of dealing with critical mass problems in collaborative systems.
Future Work

- **Scalability**
  - Efficient $k$-nearest neighbors algorithms exist
- **Interfaces that enable the user to have accurate expectations about the automated features of the system**
  - Where will my calling card be seen?
  - Interfaces for introduction
  - Exposing more internal state so people can \texttt{`debug`} inappropriate associations
- **Semantics of good collaborators (due to Larry Birnbaum)**
- **The Semantic Web – moving up to higher level representations**
The Vision – Frictionless Information Systems

- As you work the resources you need are delivered to you automatically
  - People (I2I)
  - Documents (Watson)
- No Queries, Sites, Rooms, or Places
- Instead, personalized, contextually-relevant content
  - A system aware of your goals will dynamically gather resources from relevant sources on your behalf
  - You will be free to pursue your goals instead of getting hung up on instrumental tasks
More Information

- [http://infolab.northwestern.edu](http://infolab.northwestern.edu)
- budzik@cs.northwestern.edu