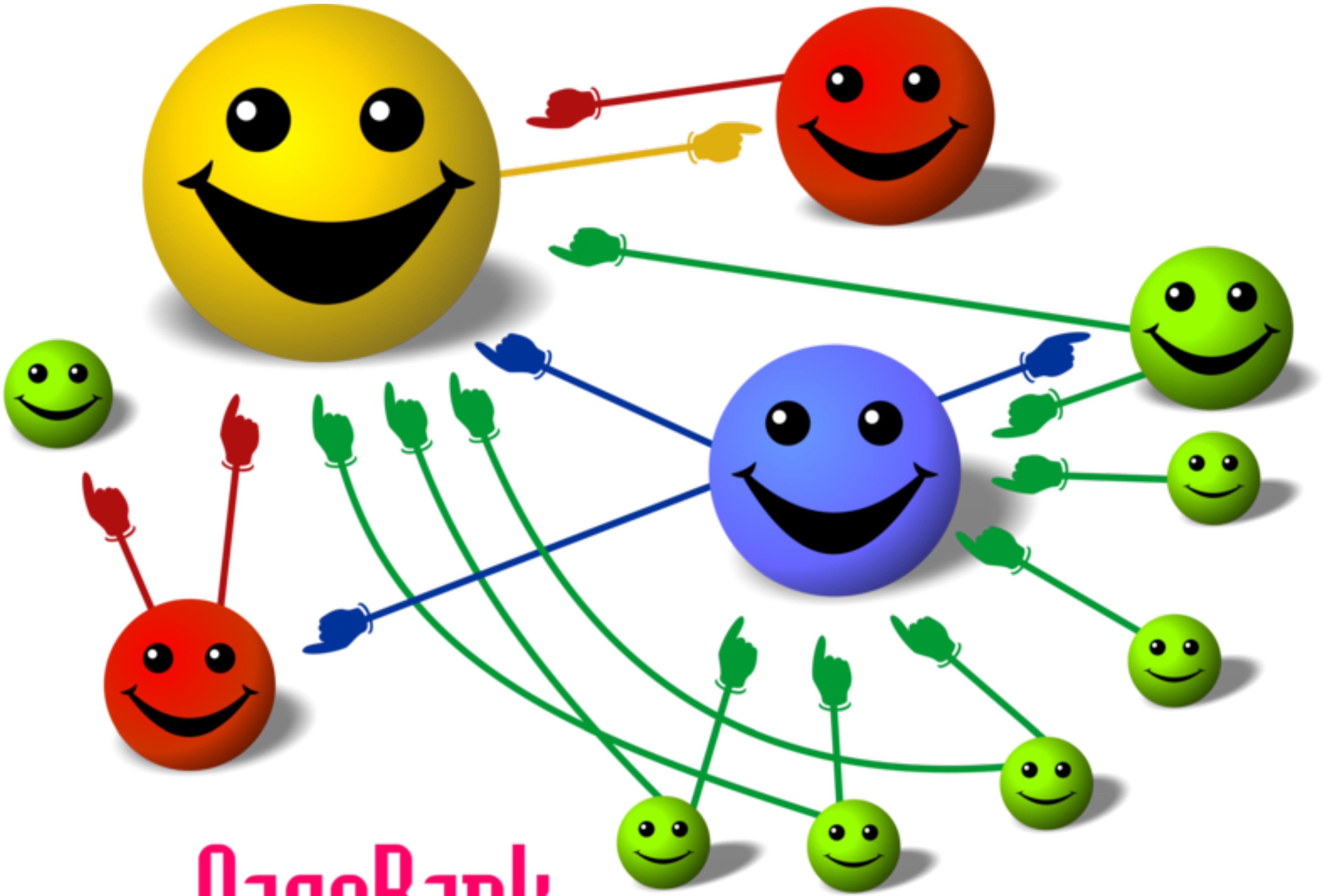
Web Search
Information Retrieval

“Web Data Analytics”

- Get the digital data (from web or from scanning)
  - Need to crawl web (? Solved “engineering” problem)
- Preprocess data to get searchable things (words positions)
- Form **Inverted Index** mapping words to documents
- Typically use **TF-IDF** (term frequency, Inverse Document frequency) to quantify importance of word match
- Rank relevance of documents: **PageRank**
- Lots of technology for advertising, “reverse engineering” “preventing reverse engineering”
- Clustering of documents into topics (as in Google News)

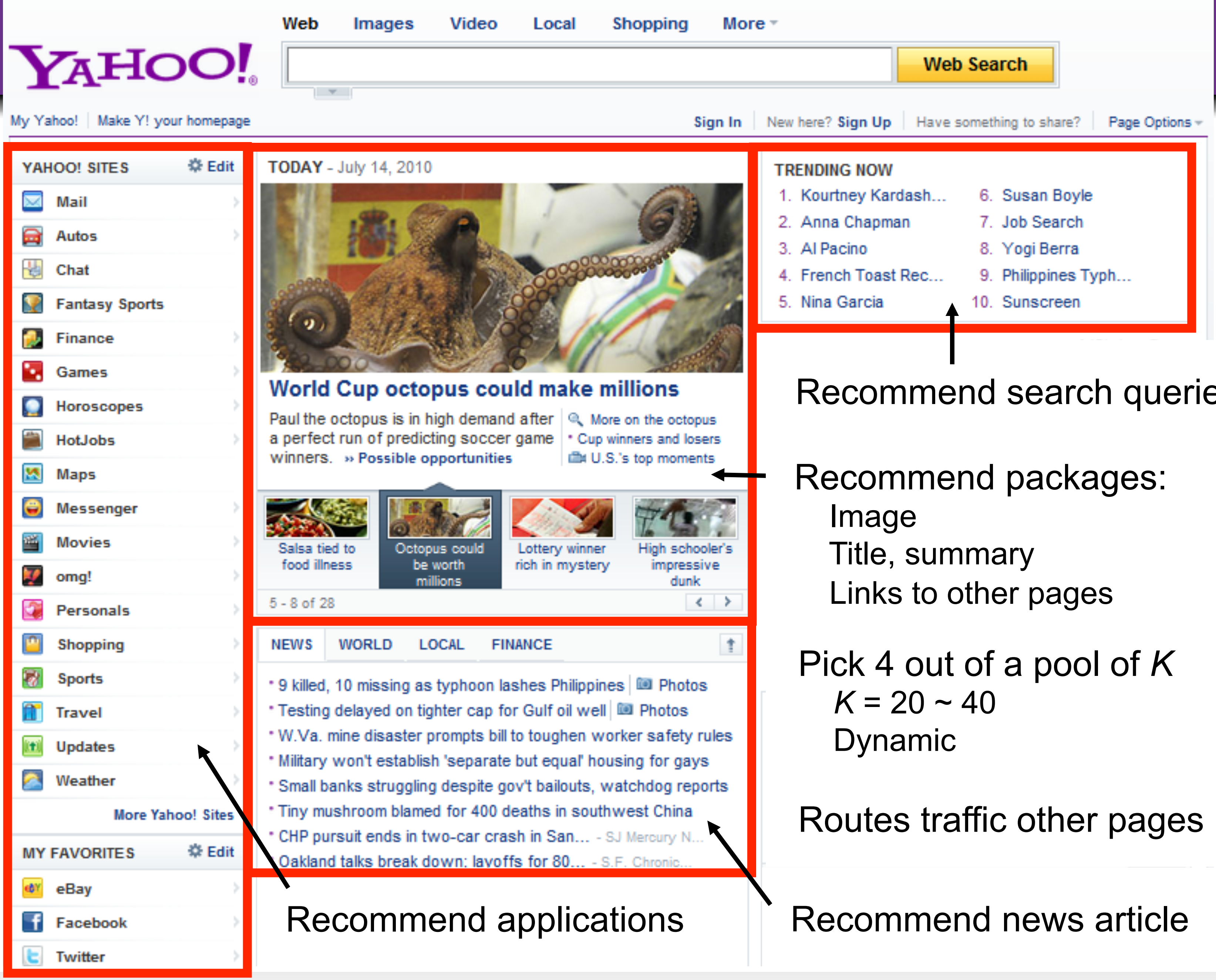


PageRank

Size of face proportional to PageRank

Modern Recommendation Systems (from Yahoo)

- Goal (Function to Optimize – Long Term dollars)
  - Serve the right item to a user in a given context to optimize long-term business objectives
- A scientific discipline that involves
  - Large scale Machine Learning & Statistics
    - Offline Models (capture global & stable characteristics)
    - Online Models (incorporates dynamic components)
    - Explore/Exploit (active and adaptive experimentation)
  - Multi-Objective Optimization
    - Click-rates (CTR), Engagement, advertising revenue, diversity, etc
  - Inferring user interest
    - Constructing User Profiles
  - Natural Language Processing to understand content
    - Topics, “aboutness”, entities, follow-up of something, breaking news,…



YAHOO!

Web | Images | Video | Local | Shopping | More

Web Search

My Yahoo! | Make Y! your homepage

Sign In | New here? Sign Up | Have something to share? | Page Options

YAHOO! SITES ⚙ Edit

- Mail
- Autos
- Chat
- Fantasy Sports
- Finance
- Games
- Horoscopes
- HotJobs
- Maps
- Messenger
- Movies
- omg!
- Personals
- Shopping
- Sports
- Travel
- Updates
- Weather

More Yahoo! Sites

MY FAVORITES ⚙ Edit

- eBay
- Facebook
- Twitter

TODAY - July 14, 2010

World Cup octopus could make millions

Paul the octopus is in high demand after a perfect run of predicting soccer game winners. » Possible opportunities

- More on the octopus
- Cup winners and losers
- U.S.'s top moments

Salsa tied to food illness | Octopus could be worth millions | Lottery winner rich in mystery | High schooler's impressive dunk

5 - 8 of 28

TRENDING NOW

1. Kourtney Kardash...
2. Anna Chapman
3. Al Pacino
4. French Toast Rec...
5. Nina Garcia
6. Susan Boyle
7. Job Search
8. Yogi Berra
9. Philippines Typh...
10. Sunscreen

NEWS | WORLD | LOCAL | FINANCE

- 9 killed, 10 missing as typhoon lashes Philippines | Photos
- Testing delayed on tighter cap for Gulf oil well | Photos
- W.Va. mine disaster prompts bill to toughen worker safety rules
- Military won't establish 'separate but equal' housing for gays
- Small banks struggling despite gov't bailouts, watchdog reports
- Tiny mushroom blamed for 400 deaths in southwest China
- CHP pursuit ends in two-car crash in San... - SJ Mercury N...
- Oakland talks break down: layoffs for 80... - S.F. Chronic...

Recommend search queries

Recommend packages:
  Image
  Title, summary
  Links to other pages

Pick 4 out of a pool of K
  K = 20 ~ 40
  Dynamic

Routes traffic other pages

Recommend applications

Recommend news article

Some examples from content optimization

- Simple version
  - I have a content module on my page, content inventory is obtained from a third party source which is further refined through editorial oversight. Can I algorithmically recommend content on this module? I want to improve overall click-rate (CTR) on this module
- More advanced
  - I got X% lift in CTR. But I have additional information on other downstream utilities (e.g. advertising revenue). Can I increase downstream utility without losing too many clicks?
- Highly advanced
  - There are multiple modules running on my webpage. How do I perform a simultaneous optimization?