# Weblog recommendation using association rules

J.J. Merelo, P. Castillo, B. Prieto, J. Carpio, F. Tricas, G. Ferreres

Dpto. de Informática e Ingeniería de Sistemas del Centro Politécnico Superior.
Universidad de Zaragoza, Spain
http://www.cps.unizar.es/~ftricas/
ftricas@unizar.es

Web Based Communities 2006. San Sebastián. Spain. February, 26-28

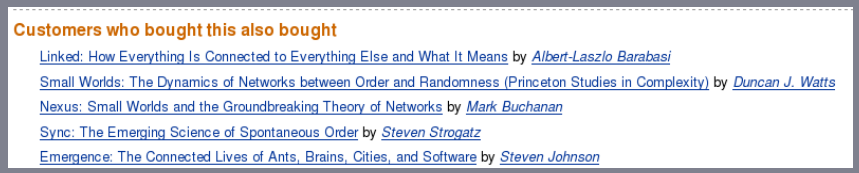
# Contents

# Objective of this paper

- We would like to recommend other blogs to weblog readers.

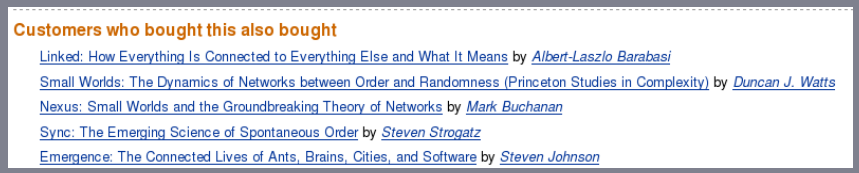
# Objective of this paper

- We would like to recommend other blogs to weblog readers.
- Something similar to what others are doing (for example, Amazon and books)



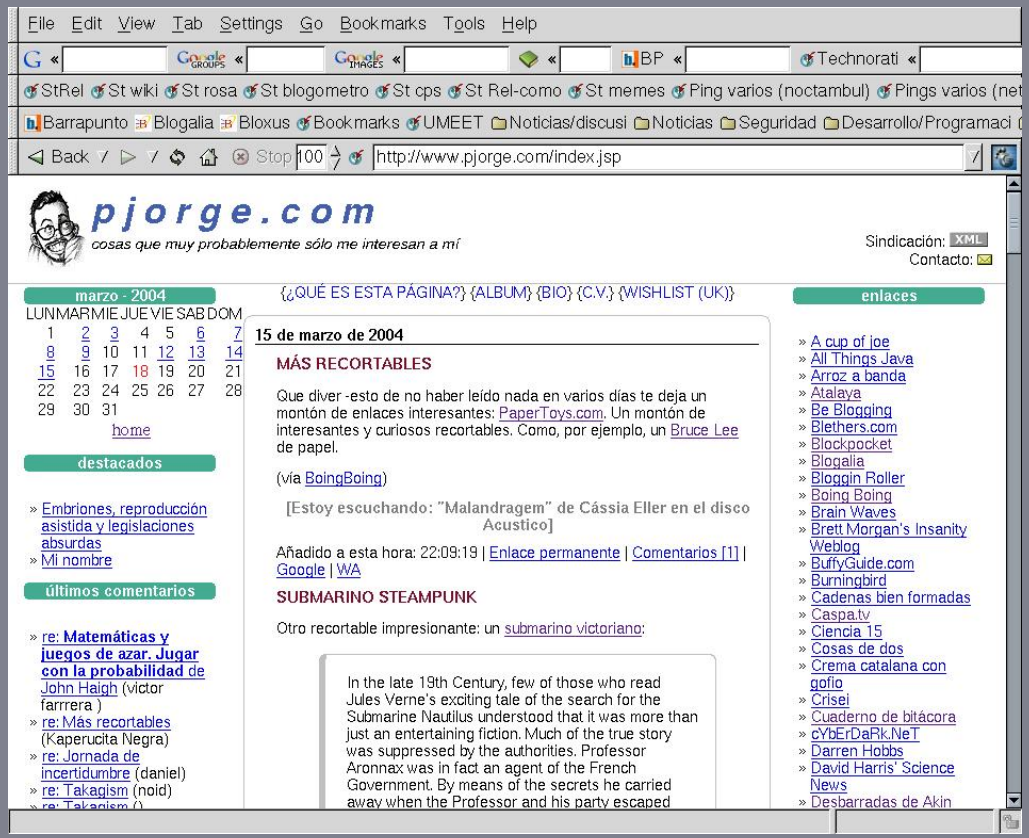
# Objective of this paper

- We would like to recommend other blogs to weblog readers.
- Something similar to what others are doing (for example, Amazon and books)




**Our recommendation:**
"People who read this blog, perhaps would like this other ones"

# What is a blog?




- Reverse-chronological order. Comments.
- Alternative formats (Rich Site Summary, RSS).
- Other stuff: polls, blogroll, music, books, latest comments...

# Blog features

- Fast growth (and growing)
  - Around 25 million acording to Technorati, PubSub, BlogPulse
  - Doubling its size about every 5-6 months (`http://www.sifry.com/alerts/archives/000419.html`)
  - Changing every day.

# So the problem is . . . ?

- Word based systems (Google, Yahoo, Altavista, . . . ) don't work well because of lack of semantics and speed.
- Specialized tools (Technorati, PubSub, BlogPulse, . . . ) solve the speed problem, but not the others (tags can help)
- Link based tools help but not every people links to the same sites, even being interested in the same things.

# We had some data...

- First Spanish webloggers and blog readers poll.
  ('I Encuesta a webloggers y lectores de blogs'
  `http://tintachina.com/archivo/cat_i_encuesta_webloggers.php`)
  - Gemma Ferreres, Antonio Cambronero
  - Self–administered
  - May, 31, June, 18 2004
  - 1662 replied (1125 bloggers, 537 readers)
  - Some interesting findings
  - Questions about blog reading

# We had some data...

- First Spanish webloggers and blog readers poll.
  ('I Encuesta a webloggers y lectores de blogs'
  `http://tintachina.com/archivo/cat_i_encuesta_webloggers.php`)
  - Gemma Ferreres, Antonio Cambronero
  - Self–administered
  - May, 31, June, 18 2004
  - 1662 replied (1125 bloggers, 537 readers)
  - Some interesting findings
  - Questions about blog reading
  - There was a second edition, last year (almost doubled participation)

# Data mining to the rescue

- Process of extraction of knowledge from huge amounts of data [DataMining, Concepts and Techniques]
- In this case, extraction of association rules:
  - Recommendation of weblogs from sets of weblogs read by users.

# Association rules I

- Composed of:

$$\text{Antecedent} \rightarrow \text{Consecuent}$$

- CD Burner $\rightarrow$ Blank CDs
  - Support 10 % (10 % transactions contain antecedent and consequent)
  - Confidence 70 % (70 % of transactions that contain antecedent, contain consequent).

# Association rules II

A Priori algorithm.

- Developed by Agrawal to analyze user purchase intentions in supermarkets (market basket analysis).
- From a database of supermarket baskets, or 'itemsets', a set of association rules that predict purchase patterns can be extracted
  - If you buy a computer, you'll want a subscription to a computer mag(other did it before you)

# Some work . . .

- Data cleaning (noise, inconsistent data, . . . )
- Data integration (combination of several sources)
- Data selection (more relevant ones)
- Data transformation (adequate format)
- Data minig (pattern extraction)
- Pattern evaluation (Obtaining the interesting rules)
- Knowledge representation

# Results analysis

- Low support due, mainly, to the excesive offer (URLs)
  Support: 0,001 to obtain 96 rules.
- Anyway, checking by hand the results they seem to be interesting and useful
- Some relations difficult to find with other techniques are detected
  Example:
  dtremmnes.net (women related), chicle.bublegum.net (a woman's fotolog) $\rightarrow$ minid.net (a guy posting about design and technology).

# Conclusions and future work

- Some useful results with a small dataset (1473 replies analyzed).
- Interesting relations emerged, that would be difficult to detect.

# Conclusions and future work

- Some useful results with a small dataset (1473 replies analyzed).
- Interesting relations emerged, that would be difficult to detect.

The future . . .

- Another types of recommendations?
  - Posts related to the ones in your blog
  - Authors with similar interests
- Improve algorithm (more efficiency, methods for obtaining better min confidence, . . . ).

# Thank you!

¿Questions?