

Weblog recommendation using association rules

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

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
- About blogs
- The problem
- Our dataset
- Data mining
- Results
- Conclusions and future work

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)

Customers who bought this also bought

[Linked: How Everything Is Connected to Everything Else and What It Means](#) by [Albert-László Barabási](#)

[Small Worlds: The Dynamics of Networks between Order and Randomness \(Princeton Studies in Complexity\)](#) by [Duncan J. Watts](#)

[Hexus: Small Worlds and the Groundbreaking Theory of Networks](#) by [Mark Buchanan](#)

[Sync: The Emerging Science of Spontaneous Order](#) by [Steven Strogatz](#)

[Emergence: The Connected Lives of Ants, Brains, Cities, and Software](#) by [Steven Johnson](#)

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)

Customers who bought this also bought

[Linked: How Everything Is Connected to Everything Else and What It Means](#) by [Albert-László Barabási](#)

[Small Worlds: The Dynamics of Networks between Order and Randomness](#) (Princeton Studies in Complexity) by [Duncan J. Watts](#)

[Hexus: Small Worlds and the Groundbreaking Theory of Networks](#) by [Mark Buchanan](#)

[Sync: The Emerging Science of Spontaneous Order](#) by [Steven Strogatz](#)

[Emergence: The Connected Lives of Ants, Brains, Cities, and Software](#) by [Steven Johnson](#)

Our recommendation:

“People who read this blog, perhaps would like this other ones”

What is a blog?

The screenshot shows the homepage of **pjorge.com**. At the top, there's a navigation bar with 'File Edit View Tab Settings Go Bookmarks Tools Help'. Below it is a search bar and a list of links. The main content area features a calendar for March 2004, a 'MÁS RECORTABLES' section with a featured article titled 'SUBMARINO STEAMPUNK', and a sidebar with 'enlaces' and 'últimos comentarios'.

CALENDARIO
LUNAMARMIÉ JUEVE SABDOM
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
29 30 31
[home](#)
destacados
» [Embriones, reproducción asistida y legislaciones absurdas](#)
» [Mi nombre](#)
últimos comentarios
» [re: Matemáticas y juegos de azar, jugar con la probabilidad de John Nash](#) (vicior farrera.)
» [re: Más reportajes](#) (Kaperucha Negra)
» [re: Jornada de incertidumbre](#) (daniel)
» [re: Talagism \(noid\)](#)
» [re: Talagism \(1\)](#)

¿QUÉ ES ESTA PAGINA? (ALBUM) (BIO) (CV) (WISHLIST (UK))
15 de marzo de 2004
MÁS RECORTABLES
Que diver -esto de no haber leído nada en varios días te deja un montón de enlaces interesantes: [Paper Toys.com](#). Un montón de interesantes y curiosos recortables. Como, por ejemplo, un [Brace Lee](#) de papel.
([vía BeingBoing](#))
[Estoy escuchando: "Malandragem" de Cassia Eller en el disco Acústico]
Añadido a esta hora: 22:05:13 | [Enlace permanente](#) | [Comentarios \(1\)](#) | [Google](#) | [WA](#)
SUBMARINO STEAMPUNK
Otra recortable impresionante: un [submarino victoriano](#)
In the late 19th Century, few of those who read Jules Verne's exciting tale of the search for the Submarine Nautilus understood that it was more than just an entertaining fiction. Much of the true story was suppressed by the authorities. Professor Aronax was in fact an agent of the French Government. By means of the secrets he carried away when the Professor and his party escaped

enlaces
» [A cup of Joe](#)
» [All Things Java](#)
» [Amoz a banda](#)
» [Atalaya](#)
» [En Blogging](#)
» [Bethers.com](#)
» [Block-pocket](#)
» [Bicajala](#)
» [Bloggin' Potter](#)
» [Boing Boing](#)
» [Brain Waves](#)
» [Brent Morgan's Inanity](#)
» [Burningbird](#)
» [Burningbird](#)
» [Cadenas bien formadas](#)
» [Caspas.tv](#)
» [Ciencia 15](#)
» [Cosas de dos](#)
» [Crema catalana con apilo](#)
» [Crisis](#)
» [Cuaderno de bitácora](#)
» [Cofre Darki.net](#)
» [Darren Hobbs](#)
» [David Harris' Science News](#)
» [Deharradas de Akin](#)

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

- Fast growth (and growing)
 - Around 25 million according 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:

Antecedent \rightarrow Consequent

- 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 mining (pattern extraction)
- Pattern evaluation (Obtaining the interesting rules)
- Knowledge representation

Results analysis

- Low support due, mainly, to the excessive 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.buble gum.net (a woman's fotolog) → 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?