



# Context Representation for Web Search Results

Jesús Vegas  
Department of Computer Science  
U. Valladolid



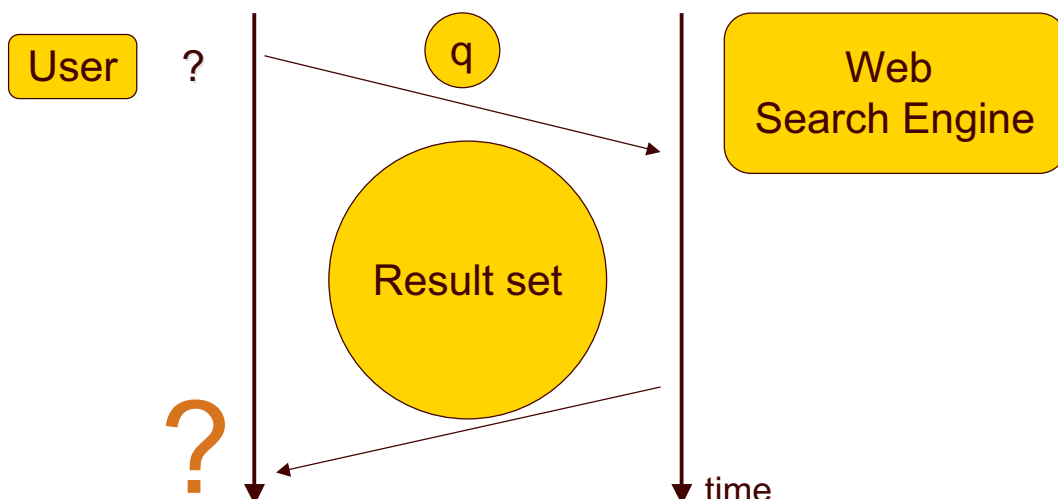
## Outline

- ➔ Intro
- ➔ Web search results in the web site context
- ➔ Visualizing complex information
- ➔ The webdocball
- ➔ The Bubble system
- ➔ Evaluation
- ➔ Conclusions and Future work

# Introduction

- ➔ Searching the web is one of the most frequent information access tasks nowadays
- ➔ ... but it is often also one of the most frustrating ones.
- ➔ More efforts designing and developing algorithms to improve the query process, less done to help the users to analyze the results.

# Searching the Web



# Searching the Web

- ➔ Search engine research focused with the search process (efficiency and effectiveness).
- ➔ It's up to the user to determine the page to get.

Visualization

to find the needle in the document haystack

# Web Result Visualization

- ➔ Keyword in Context (KWIC)

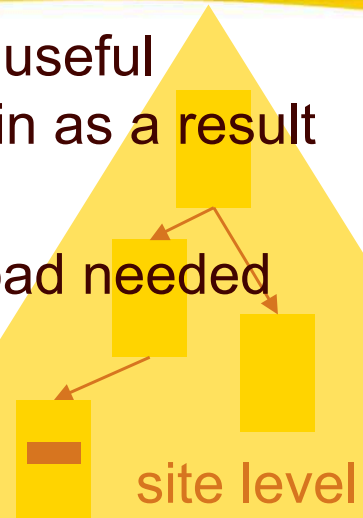
The screenshot shows a Yahoo! search result for the query 'web result visualization'. The search bar contains the text 'web result visualization' and a 'Buscar' button. Below the search bar, there are navigation links for 'Toda la Web', 'Imágenes', 'Video', 'Directorio', 'Pág. Amarillas', 'Noticias', 'Compras', and 'Más »'. The search results section shows 'Resultados de búsqueda' with '1 - 10 de aproximadamente 1.620.000 de web result visualization 0,23 seg.'. The first result is a PDF document titled 'WebQuery: Searching and Visualizing the Web Through Connectivity (PDF)'. The snippet for this result reads: '... results of a Web query, it also extends the result set beyond what ... Information visualization, World-Wide Web, Information Retrieval ... Result ... cybermetrics.cindoc.csic.es/cybermetrics/pdf/24.pdf - 100k - Ver como html - Más resultados de este sitio web'. To the right of the search results, there is a Google search bar with the text 'web result visualization' and a 'Búsqueda' button. Below the Google search bar, there is a 'La Web' section with a search bar and a 'Resu' button. The snippet for this result reads: 'Formato de archivo: PDF/Adobe Acrobat Multifom Glyph Based Web Search Result Visualization. Jonathan Roberts, Nadia Boukheiffa, Peter Rodgers. University of Kent at Canterbury ... ieexplora.ieee.org/ie5/7989/22105/01028828.pdf - Páginas similares'.

- ➔ It fails to show how a relevant page is related to the content of the entire web site it belongs to.

# Designing interfaces

- ➔ to maximize the amount of useful information that users obtain as a result of a web search, and
- ➔ to minimize the cognitive load needed to interpret it.

document level



Vivísimo - Clustered search on web result visualization

http://vivisimo.com/search?tb=vivisimocom&query=v

about | products | solutions | press | partners | support

web result visualization

the Web

Search **Clusty.com** with our **Firefox Toolbar**

Advanced Search Help

Clustered Results

web result visualization (178)

Search Result Visualization (31)

Visual Bracketing for Web search Result Visualization (5)

Multiform Glyph Based (4)

Multiple Search (4)

Categorized (5)

Tools, Scientific (4)

Visual Depictions Of Search Results (2)

Web-based (3)

Other Topics (6)

Data (31)

Science (15)

Software (17)

Terminado

Top 176 results of at least 398,000 retrieved for the query web result visualization (Details)

**Affirmations** [new window] Sponsored Link  
The Power Of My Subconscious Mind, Will Create Any Reality I Choose!  
www.Secret-Biz.com - Sponsored Listings 1

**3D visualization** [new window] Sponsored Link  
High quality stereo 3D projection systems.  
www.cyviz.com - Sponsored Listings 2

1. **International Conference on Information Visualisation 2003**  
[new window] [frame] [cache] [preview] [clusters]  
**Visual Bracketing for Web search Result Visualization**. 264-271 Electronic Edition (link) BibTeX.  
Applications of Graph Theory. Jean Flower, Peter Rodgers, Paul Mutton  
www.acm.org/sigs/sigmod/dblp/db/conf/iviv2003.html - Wisenut 3, Ask 30, Ask 35, MSN 48

2. **Multiform Glyph Based Web Search Result Visualization**  
[new window] [frame] [cache] [preview] [clusters]  
Spreadsheet Validation and Analysis through Content Visualization Multiform Glyph Based Web Search Result Visualization  
www.oculusinfo.com/papers/SpshContentViz-Eusprig06-Mar31-final.pdf - MSN 1, Ask 2, Ask 3, MSN 3, Ask 4, Ask 8, Ask 11

3. **Computer Science: Applied and Interdisciplinary Informatics Group publications** [new window] [frame] [preview] [clusters]  
Analysis of a Multiple Search Result Visualization Edward 2004 Textual Difference

KartOO Metamotor de búsqueda

http://www.kartoo.com/flash04.php3

help << web result visualization Search search the web

www.kartoo.com

starlight.pnl.gov

www.cgl.uwaterloo.ca space

www-2.cs.cmu.edu interactive

information www.dcs.gla.ac.uk analysis

sdl.computer.org demonstrate searching conference

www.kgs.ku.edu multiple glyph based www.cvev.org www.asis.org

www.oculusinfo.com

help www.cs.kent.ac.uk jonathan roberts nadia boukhelifa

http://www.dcs.gla.ac.uk/~jhw/haptic.pdf

> ToileQuebec Hotbot

Context Representation for Web Search Results 9

di Departamento de Informática UVa

## Document level

- ⇒ The structure of the web page could be of great interest for the user
  - it can show the overall complexity of the page
  - and where the relevant parts of the document are.
- ⇒ This is especially important for complex pages, which use different structural elements like tables, graphics, columns, etc.



## Web site level

- ⇒ The web pages returned as results of a query could be displayed in relation to the overall structure of the web sites they belong to.
- ⇒ Web sites could be ordered by their relevance with respect to the query.



## Relevance of a web site?

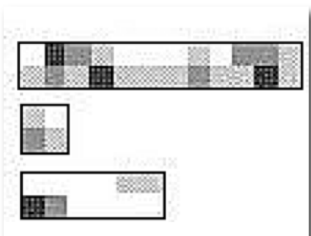
- ⇒ If many of its web pages are relevant to the query the site is relevant.
- ⇒ Other factors:
  - How deeply a page is in the site?
  - Are there other relevant web pages linked to it?
  - It can give the user an overview of the context in which the potentially relevant page is placed.
  - Browsing for finding information.

Taking the structure into account is necessary

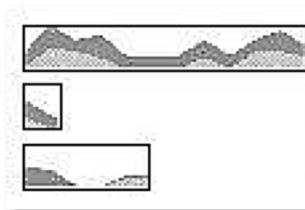
# Results context

- ➔ To reduce the cardinality of the results set without losing recall in the process.
- ➔ We use information visualization to achieve this.
- ➔ To add an additional logical level of results presentation before the web page level: the web site level.

# Tilebars, Relevance Curves and Thumbnails



➔ Tilebars

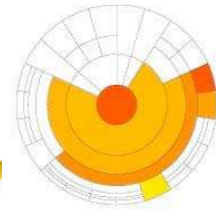


➔ Relevance Curves



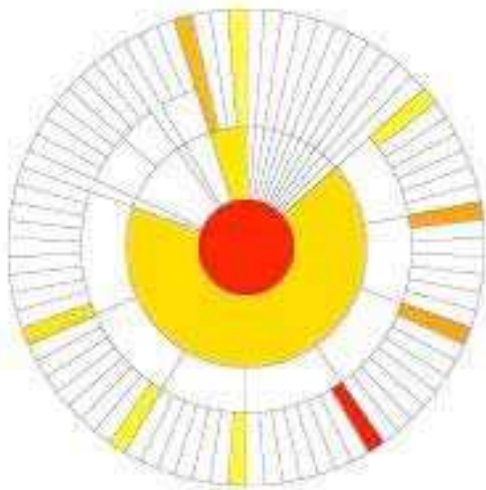
➔ Thumbnail

# DocBall



- ➔ An extension of the DocBall visualization metaphor which enables the visualization of both web pages and web sites in a clear and consistent way.
- ➔ This metaphor has been applied to the design and implementation of a proof-of-concept web search engine called Bubble.

# WebDocBall



- ➔ An extension of the DocBall metaphor to the Web domain
  - Web pages has some structure: HTML, XML.
- ➔ A visual metaphor that can explain the user
  - why a Web document was retrieved
  - by showing where in the structure of the document estimated relevance lies.

# Extracting Structure from a Web Document

## ➔ Random URL generator using Google

- 30,000 distinct HTML pages,
- 13,124,339 HTML labels
- the most often used labels were
  - content, p-IMPLIED, td, br, tr, img, p, comment, table, span, li, script, div, meta, input, title, center, hr, ul, body, html, head, form, link, select, frame, object, ol.

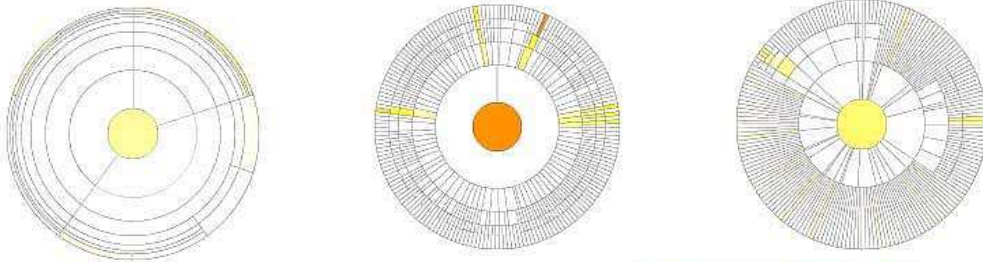
## ➔ We consider only those labels with evident structural meaning, used to express hierarchical relations.

The screenshot shows a web browser window displaying an HTML document titled "Introduction to IR". The document content includes a navigation menu with links for "Index", "Teach", "Research", and "Misc", followed by sections for "Description", "Program", "Laboratories", and "References". Each section contains specific content, such as a paragraph in "Description", a list in "Program", a link in "Laboratories", and a list of references in "References".

Overlaid on the right side of the browser window is a circular diagram that maps the HTML structure of the document. The diagram is a tree-like structure where the root is a `<table>` tag. This root branches into several `<td>` tags. Some of these `<td>` tags further branch into other tags, such as `<table>`, `<table>`, `<ul>`, `<li>`, `<p>`, `<ol>`, and `<li>`. The diagram visually represents the nested and sequential relationships between the HTML elements in the document.

# WebDocBall

- ➔ Can be considered as the representation of a tree structure.
- ➔ Web documents with many levels, the height of the rings decreases the deeper the structural level it represents.



# DocBall applied to a Web Site

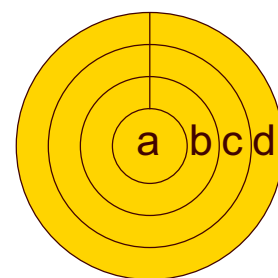
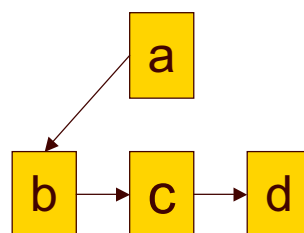
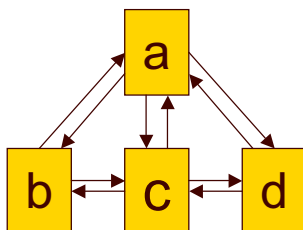
- ➔ When we apply the DocBall representation to web sites the objective is to show to the user:
  - the structure of the web site,
  - which are the relevant pages in the site,
  - and how relevant is the whole site.
- ➔ This allows the user to have an overall view of the site but also to go directly to the most relevant pages in the site.

# From a web site to a WebSiteBall

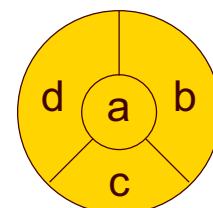
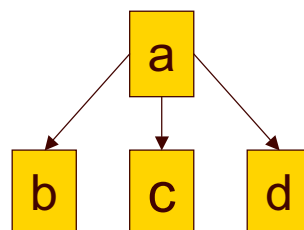
- ➔ Some general hypotheses for most web sites:
  - Web sites are relatively homogeneous sets of web pages, that is, they are about loosely related topics. So linked web pages are somehow topically related.
  - Pages of the web sites are linked with a large number of redundant links, for the user's convenience.
- ➔ We need to transform the web site traversal graph into a tree.

# From a web site to a WebSiteBall

➔ deep-first



➔ breadth-first



## From a web site to a WebSiteBall

- ➔ In the first prototype, we used a very simple propagation function:
  - each page's relevance is evaluated independently,
  - but just computes it for the entire site using the average relevance value of all web pages in a site.

$$q_i = \begin{cases} f_{\text{relevance}}(s_i) & l_i \neq 0 \\ \frac{\sum_{i=1}^n q_i}{n} & \text{otherwise} \end{cases}$$

## The Bubble System

- ➔ Web search engine called Bubble that uses WebSiteBalls and WebDocBalls to represent the results of a search.
- ➔ The user can choose among three different query objectives:
  - web sites;
  - web pages;
  - specific file types in web pages.

The screenshot shows the Buble search engine interface. At the top, there is a search bar with the text 'Vegas' and a 'Search' button. Below the search bar, there are three radio buttons: 'Site Web Search' (selected), 'Page Web Search', and 'Web Site Structure'. To the right of these buttons is a 'Navigate' button. The search results are displayed in a list format, with each result showing a title, a URL, and a score. To the right of each result is a circular context representation chart. The charts are colored in shades of green and yellow, indicating the relevance of different terms in the search results.

Searched the web for **Vegas** Results 0 - 5 of about 5

**Página personal de Joaquín Adiego**  
<http://www.infor.uva.es/~jadiego> Score: 6  
[http://www.infor.uva.es/~jadiego/files/ej\\_prog\\_1.pdf](http://www.infor.uva.es/~jadiego/files/ej_prog_1.pdf)

**Página principal**  
<http://www.infor.uva.es/~julian> Score: 2  
<http://www.infor.uva.es/~julian/investiga.html>

**Miguel Angel Villarroel Salgueiro's Home Page**  
<http://www.infor.uva.es/~miguelv> Score: 1  
<http://www.infor.uva.es/~miguelv/lapaz.html>

Context Representation for Web Search Results

The screenshot shows the Buble search engine interface with search results for 'Vegas'. The search bar contains 'Vegas' and the 'Search' button is visible. The search results are displayed in a list format, with each result showing a title, a URL, and a score. To the right of each result is a circular context representation chart. The charts are colored in shades of blue and cyan, indicating the relevance of different terms in the search results. Below the search results, there is a bubble chart with the text 'bub b b b b b b b b b le' and numbers 1 through 10 below it.

Searched the web for **Vegas** Results 0 - 5 of about 100

**1) Investigación**  
<http://www.infor.uva.es/~jadiego/invest.htm> Score: 6  
 Tag: < TEXTO >  
 Valor: , vol. 20, n 4, 2002. ISSN 0264-0473, paginas 306-313.

**2) Programa ABD**  
<http://www.infor.uva.es/~julian/progr-ip.html> Score: 2  
 Tag: < TEXTO >  
 Valor: Se plantearan ejercicios practicos en cada bloque de temas que se recogeran y corregiran. Examen de

**3) Miguel Angel Villarroel Salgueiro's Home Page**  
<http://www.infor.uva.es/~miguelv/resumesp> Score: 1  
 Tag: < TEXTO >  
 Valor: Alberto Pedrero, Pablo de la Fuente, Miguel Villarroel y Jesus Vegas,

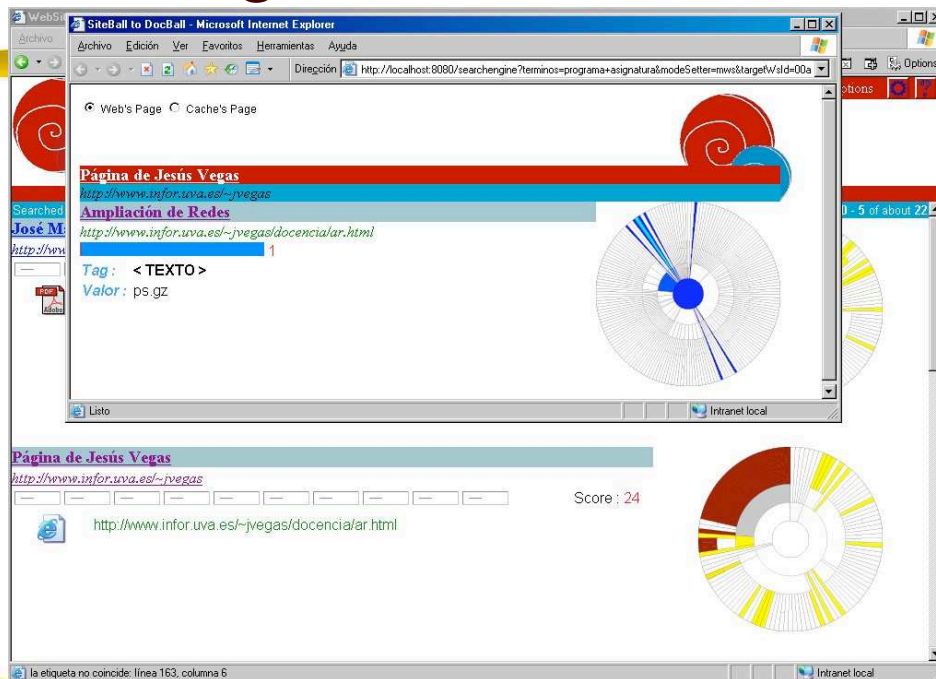
**4) Miguel Angel Villarroel Salgueiro's Home Page**  
<http://www.infor.uva.es/~miguelv/resume.htm> Score: 1  
 Tag: < TEXTO >  
 Valor: The thesis presents a user interface design method that applies the Object Oriented Model. In first term is

**5) Miembros**  
<http://www.infor.uva.es/personal.htm> Score: 1  
 Tag: < b >  
 Valor:

◀ bub b b b b b b b b b le ▶  
 1 2 3 4 5 6 7 8 9 10

Context

# Viewing a WDB from a WSB



Context Representation for Web Search Results

27

# Evaluation

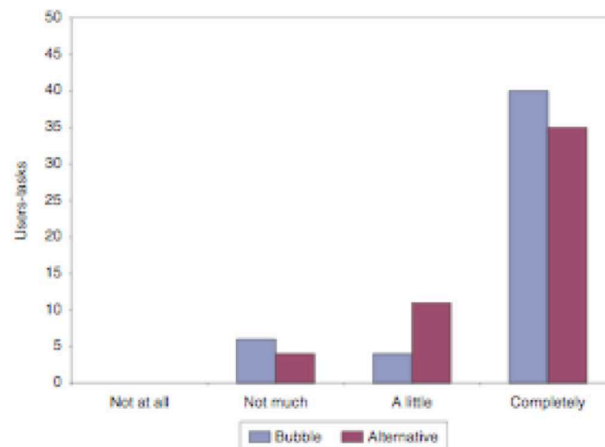
- ⇒ Using a user and task oriented approach
  - the full site of the University of Strathclyde,
  - 10 users to locate pages relevant to some user-defined task
  - bubble and alternative system that presented the results as a flat list of web pages
  - both with the same model to estimate relevance

Context Representation for Web Search Results

28

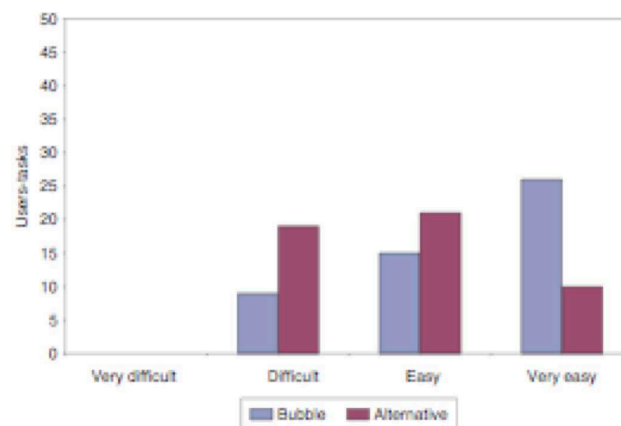
# Evaluation

⇒ Q1: Rate how much you were able to accomplish the task



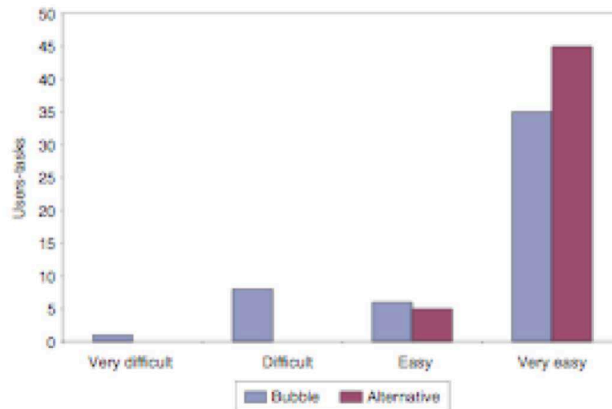
# Evaluation

⇒ Q2: Rate the complexity of the task completion



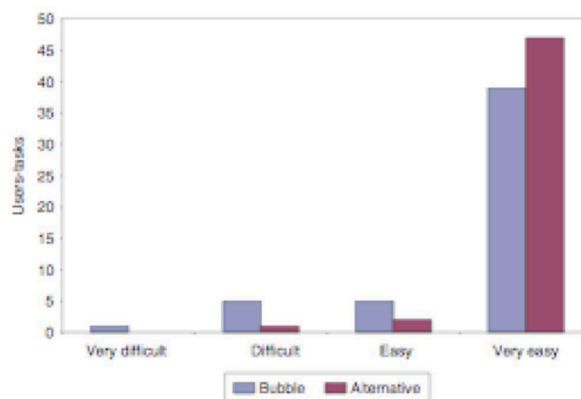
# Evaluation

⇒ Q3: Rate the complexity of the use of the system for this task



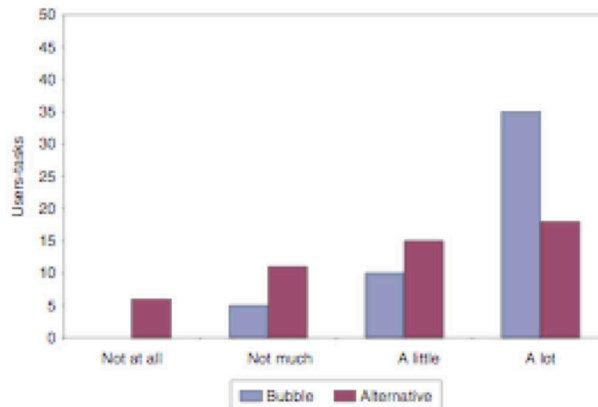
# Evaluation

⇒ Q4: Rate the complexity of the interpretation of the search results for this task



# Evaluation

- ➔ Q5: Rate how the results presentation helped you find the information sought for this task



# Conclusions

- ➔ It is very important to design and develop information visualization techniques that represent the context in which a document is estimated to be relevant.
- ➔ This is particularly important in web searching, given the complexity of the task and the complexity and ambiguity of the relations existing between web pages in a web site and between elements of a web page.

# Future work

- Redesign of Bubble, following a formative design approach.
- To design of a more powerful relevance propagation algorithm that should enable a more precise and fast estimate of the relevance of a web site from its page

# Context representation for web search results

- Questions?

