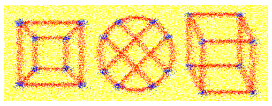

Journal of Graph Algorithms and Applications
http://jgaa.info/ vol. 15, no. 2, pp. 229–251 (2011)

A Graph Drawing Application to Web Site Traffic Analysis

Walter Didimo[1] *Giuseppe Liotta*[1] *Salvatore A. Romeo*[1]

[1]Dipartimento di Ingegneria Elettronica e dell'Informazione
Università degli Studi di Perugia, Italy

Abstract

Web site traffic analysis studies how different pages forming Web sites are accessed by their target audience over time. From a business intelligence point of view, it represents one of the key activities for many private companies and public institutions. Recent papers pointed out that Web site traffic analysis is especially useful if it focuses on the user interest into the relevant *concepts* described in a Web site rather than counting user accesses to the distinct *pages* forming the Web site. This paper extends existing measures of conceptual Web site traffic analysis and describes a new system that supports this analysis by means of graph visualization techniques. The graph drawing engine of the system is a force-directed heuristic that computes a simultaneous embedding of two suitably defined graphs, that are non-planar in general. The heuristic is especially designed to take into account two important aesthetic criteria of the drawing: crossing resolution and geodesic edge tendency. We also present some experiments and case studies to show the effectiveness of the proposed approach in practice.

Submitted:	Reviewed:	Revised:	Accepted:
June 2010	October 2010	December 2010	March 2011
	Final:	Published:	
	March 2011	March 2011	
Article type:		Communicated by:	
Regular paper		I.G. Tollis	

The research in this paper has been partially supported by the Umbria Region under the FESR grant "COWA: A Conceptual Web Analyzer" and by MIUR of Italy under project AlgoDEEP Prot.2008TFBWL4. Since the grant is related to an industrial project, we cannot distribute the system. Also the real data sources used for testing the system cannot be revealed (as specified in the paper). However, in order to witness that the system is really working, we have produced a movie. It can be downloaded at the following URL: http://gdv.diei.unipg.it/docs/cowa/. An extended abstract of this research titled "Graph Visualization Techniques for Conceptual Web-traffic Analysis" appears in the proceedings of the 3rd IEEE Pacific Visualization Symposium (PacificVis 2010).

E-mail addresses: didimo@diei.unipg.it (Walter Didimo) liotta@diei.unipg.it (Giuseppe Liotta) romeo@diei.unipg.it (Salvatore A. Romeo)

1 Introduction

Systems for Web site traffic analysis are used by private companies and public institutions to understand how their on-line communication channels are accessed by their target audience over time. Typically, these systems analyze the data stored in Web server's log files and output different statistic reports, such as the number of accessed pages, the rank of the most accessed pages, the number of visitors in a given time frame. The reports are commonly presented as lists, tables, plots, or charts. Examples of popular systems for Web site traffic analysis are 123LogAnalyzer, Analog, AWStats, Google Analytics. The statistics produced by these systems should support business analysts in: (*i*) better understanding what items or services are the most interesting for the users, which may lead to new strategic decisions and market investments; (*ii*) discovering unexpected behaviors in the exploration of a Web site, which can give hints on how to improve the company on-line communication channel. A list of references about Web content and usage mining includes [2, 10, 24, 28, 30].

1.1 Concept-based Web Site Traffic Analysis

There is a growing consensus that classical Web site traffic analysis approaches are not fully satisfactory for current business intelligence applications. This because the statistical reports they produce, which show for instance the number of user accesses to the different *pages* of a Web site, are mainly useful for Web designers, while they may not be meaningful for business analysts. Instead, a business analyst would like to know what are the *concepts* (i.e., topics) of major interest for the visitors of the Web site. The limits of reporting the number of accesses to each Web page is even more evident if one considers that a Web page is typically referenced by its URL (Uniform Resource Locator), and the URL may be far from reflecting the page content.

An emerging research direction in Web site traffic analysis aims at decoupling the physical organization of Web sites from the concepts that they contain in a given time frame. To face this problem, Norguet *et al.* define a *concept-based audience metric* [26, 27]; this metric uses classical information retrieval techniques to extract relevant terms from the pages of a Web site and applies ontologies to recursively aggregate terms into concepts [1, 17]. The concept-based audience metric combines the volume of visitors of the different Web pages with the presence of each concept in these pages, to obtain a measure of the interest of the visitors in the different concepts. However, the presentation of the values of user interest for the main concepts in the Web pages is still based on classical graphical reports (e.g., plots and bar charts).

1.2 Our Contribution

This paper extends the concept-based audience metric model and describes a system that supports the conceptual Web site traffic analysis. While the model by Norguet *et al.* considers all concepts to be independent, our main idea is to

analyze the interaction between the different concepts and how this interaction correlates with the user interest. Our techniques allow a business analyst to: (a) estimate to what extent the interest for a concept C is affected by the user interest into other concepts that appear in the same Web pages as C; (b) discover possible relationships between concepts by revealing whether pairs of concepts are visited by a relevant amount of users within a given time frame. The intrinsic relational nature of the concepts is explored by means of an enhanced diagrammatic interface that shows multiple visualizations of graphs. In more detail, our contribution is as follows.

- We refine and generalize the concept-based metric by introducing two kinds of graphs, G_{pag} and G_{usr} (Section 2). These graphs have the same vertex set, each vertex representing a concept in a given Web site. In G_{pag} two concepts are connected if they appear simultaneously in a page of the Web site. In G_{usr} two concepts are connected if they were accessed by a user in the same browsing session.

- We describe graph visualization techniques that help the analyst to discover relevant concepts for the users and relevant correlations between these concepts (Section 3). The interaction paradigm is based on the visual comparison of G_{pag} and G_{usr}. In addition, the analyst can interact with a third graph, G_{vis}, that summarizes the distribution of the users and their browsing behavior on the visited pages.

- In order to support the visual comparison of G_{pag} and G_{usr}, we study the *simultaneous embedding problem* for these two types of graphs (Section 4). We remark that the simultaneous embeddability problem is a well known topic in the graph drawing literature (see, e.g., [3, 16]). Differently from previous work on this subject, our main goal is to optimize some aesthetic criteria that are recently shown to be of great impact on the readability of a layout. Namely, according to the cognitive experiments described in [20, 22], drawings with edge crossing angles close to 90 degrees are preferred. Additionally, when the edges are represented as simple curves, users tend to better recognize an edge if it resembles the straight segment between its end-points [21]; we call this user tendency the *geodesic edge tendency*. We present a force-directed heuristic that computes a simultaneous embedding of two non-planar graphs, and we experimentally show that it exhibits a good behavior in terms of the aesthetic criteria mentioned above.

- We describe a system, called COWA, that implements the concept-based metrics and the graph drawing techniques listed above (Section 5). We discuss case studies for the conceptual analysis of two popular Web sites with COWA, the first containing information about volleyball championship and teams, and the second containing national and international news.

We remark that our methodology does not directly compare with statistical analysis techniques that try to automatically infer if two or more phenomena

are correlated. Our aim is to support the analyst in this task by providing her with advanced visualization tools.

1.3 Related works

Our application falls in the Web mining area and combines two main ingredients for Web site traffic analysis: The use of metrics for concept analysis and the use of graph visualization techniques.

In the field of concept analysis, there are different approaches to extract concepts from text data: Natural Language Processing (NLP) is concerned with the interaction between computers and human language. NLP is used to understand the meaning of text or to process structured data and present them in a readable way [25]. Formal Concept Analysis (FCA) is a mathematical theory to analyze data, represent knowledge and manage information [19], with several applications in the field of information retrieval. FCA allows detection of concepts in a text using predefined sets of objects and related attributes in a given context [8, 9]. Another approach defines concepts in a text as an aggregation of key-terms, based for example on the use of ontologies [26, 27]. This last approach is the only introduced so far for defining concept-based metrics related to Web-traffic analysis. Similarly to [26, 27] we still assume that concepts are an aggregation of key-terms; however, in our setting they are manually defined without necessarily using ontologies.

The use of visualization techniques for the analysis of Web mining data is motivated by the growing need for the business analysts to obtain higher level information from raw data. A system that applies different visual metaphors for representing the structure of a Web site where Web metrics are overlaid is described in [29]. Systems and visual paradigms for correlating Web navigational data with the structure of Web sites are described in [5, 6, 31]. Our information visualization approach is concerned with the simultaneous display of "networks of concepts", which allows for their visual comparison. We discuss in details the related work on graph visualization algorithms in Section 4.2.

2 Concept-Based Audience Metrics and Graphs

Let $\{C_1, C_2, \ldots, C_n\}$ be a set of *concepts* addressed by a Web site. We assume that each concept C_i ($i \in \{1, 2, \ldots, n\}$) is defined as a collection of *terms* $\{s_{i,1}, s_{i,2}, \ldots, s_{i,k_i}\}$. We also say that $s_{i,j}$ *is a term of* C_i. Terms and concepts can be computed automatically, by combining text analysis techniques and ontologies [27], or they can be manually defined. When a term is a word of some dictionary, it can be assumed in its stemmed version (which depends on the chosen language).

In this paper, we assume that a set of concepts is manually defined by an analyst (e.g., a manager) who knows the contents of the Web site of her company, but who may ignore the structure of the Web site and the distribution of the contents in the different Web pages. In our scenario, the goal of the analyst is

to analyze the interest of the target audience of the Web site for the different concepts in a certain period of time.

2.1 Measuring the Interest for a Concept

In this section we give a formal definition of *interest* for a concept or for a pair of concepts. Our notions refine and generalize those by Norguet *et al.* [27]; they will be used to measure the relevance of the different concepts contained in a Web site from the user's perspective, and to help the analyst to infer possible correlations between the relevance of two distinct concepts. In this paper, the notion of *interest* introduced by Norguet *et al.* is renamed as the *gross interest* for the concept; additionally, we define the new notion of *net interest* for a concept with respect to another one.

From now on with the term *page* we refer to the content accessible at a certain URL (Uniform Resource Locator) in a given Web site. We denote by P the total set of pages of a Web site and by $P_d \subseteq P$ the subset of pages visited during day d [1]. We start by recalling the definition of two basic functions, called *Consultation* and *Presence*, introduced in the paper by Norguet *et al.*. Intuitively speaking, the Consultation measures how many times a certain term (or concept) has been accessed over a given period of time; this is done by combining the number of times that documents containing that term have been requested by the users, and the frequency of the term inside those documents. The Presence measures the "spread" of a term inside all documents over a given period of time, independent of the number of document requests. More formally, the functions are defined as follows:

- **Consultation:** For each term $s_{i,j}$ and for each page $p \in P_d$, the function $Consultation(s_{i,j}, p, d)$ represents the frequency of $s_{i,j}$ in p multiplied by the number of times page p has been displayed on visitors' screen during day d. To avoid big fluctuation over time, the consultation of a term can be divided by the total number of page accesses during d. For each concept C_i and for each page $p \in P_d$, the following function is defined: $Consultation(C_i, p, d) = \sum_{j=1}^{k_i} Consultation(s_{i,j}, p, d)$.

- **Presence:** For each term $s_{i,j}$ and for each page $p \in P$, the function $Presence(s_{i,j}, p, t)$ denotes the frequency of term $s_{i,j}$ in page p at time t. The function $Presence(s_{i,j}, p, d)$ integrates $Presence(s_{i,j}, p, t)$ over d, i.e., $Presence(s_{i,j}, p, d) = \int_d Presence(s_{i,j}, p, t)dt$. For each concept C_i and for each page $p \in P$, the following function is defined: $Presence(C_i, p, d) = \sum_{j=1}^{k_i} Presence(s_{i,j}, p, d)$. We also define the presence of a concept C_i during day d as $Presence(C_i, d) = \sum_{p \in P} Presence(C_i, p, d)$.

If $Presence(C_i, p, d) > 0$ we say that C_i *was present in* p and that p *contained*

[1] If a Web site is strongly dynamic, many URLs can depend on the user interaction, and cannot be automatically discovered in an exhaustive way. We assume as set P of pages the one obtained by all URLs contained in the Web server log file in a long period of time.

C_i, during day d. Also, if two concepts C_i and C_j are both present in the same page p during day d, we say that C_i and C_j *shared* p.

Definition 1 *The interest for a concept C_i in a subset $P'_d \subseteq P_d$ is the function*

$$Int(C_i, P'_d, d) = \frac{\sum_{p \in P'_d} Consultation(C_i, p, d)}{Presence(C_i, d)}.$$

The next definition measures the total interest for a concept C_i during a day d and it is equivalent to the definition of interest given by Norguet *et. al* [27].

Definition 2 *The gross interest for a concept C_i during day d is the function $GrossInt(C_i, d) = Int(C_i, P_d, d)$.*

Additionally, for two concepts C_i and C_j, we introduce a measure of the interest for C_i "modulo" the interest for C_j. Namely, let $P_d^{ij} \subseteq P_d$ be the subset of pages of P_d shared by C_i and C_j.

Definition 3 *The net interest for a concept C_i with respect to a concept C_j during day d is the function $NetInt(C_i, C_j, d) = GrossInt(C_i, d) - Int(C_i, P_d^{ij}, d)$.*

Notice that, $NetInt(C_i, C_j, d)$ is in general different from $NetInt(C_j, C_i, d)$.

2.2 Graphs of Concepts

Besides the concept-based metrics defined above, we introduce three types of graphs. In Section 3 we will describe a graph-drawing based interface that, by means of Definitions 1-3 and the graphs defined here, makes it possible to evaluate to what extent the interest for a concept is affected by the interest for other concepts.

The three types of graphs we use are called the *shared-pages graph*, the *shared-user graph*, and the *visited-pages graph*. They are denoted by G_{pag}, G_{usr}, and G_{vis}, respectively, and are defined as follows:

- G_{pag} and G_{usr} are defined on the same set V of vertices, i.e., $G_{pag} = (V, E_{pag})$ and $G_{usr} = (V, E_{usr})$. For each concept C_i ($i \in \{1, \ldots, n\}$), there exists an associated vertex $v_i \in V$. Let v_i and v_j be two vertices of V. There is a weighted edge $(v_i, v_j) \in E_{pag}$ if C_i and C_j shared some pages of P_d during day d; in this case, the weight of (v_i, v_j) is set equal to the number $|P_d^{ij}|$ of shared pages. There is a weighted edge $(v_i, v_j) \in E_{usr}$ if there exists a user that during day d visited some pages containing C_i and some pages containing C_j. As weight for an edge $(v_i, v_j) \in E_{usr}$ we consider two possible definitions. Namely, let $U_d^i[j]$ (resp. $U_d^j[i]$) be the set of users who visited pages of P_d containing C_i but not C_j (resp. C_j but not C_i), and let U_d^{ij} be the set of users who visited pages of P_d^{ij}. We define the following possible weights for (v_i, v_j): The *gross weight*, defined as $|(U_d^i[j] \cap U_d^j[i]) \cup U_d^{ij}|$, and the *net weight*, defined as $|(U_d^i[j] \cap U_d^j[i])|$.

- G_{vis} has a vertex u_i for each page $p_i \in P_d$, and it has a directed edge (u_i, u_j) if there is a user that visited p_j immediately after p_i during day d.

3 Graph Visualization Techniques

Given a Web site and a set of concepts $\{C_1, C_2, \ldots, C_n\}$ for this site, our main goal is to support the analyst in mining reliable information about the interest of the target audience for the different concepts in a desired period of time. Our approach is based on graph visualization and relies on comparing and interacting with graphs G_{pag}, G_{usr}, and G_{vis}. We have in mind two main tasks of analysis:

> **Task 1: Discovering relevant concepts**. The goal of this task is to understand if the gross interest for a concept is reliable or if it can be positively affected by the interest for other concepts. If in G_{pag} there are two vertices v_i and v_j that are not adjacent, we can assume that the gross interest for C_i is not affected by the interest for C_j and vice-versa. In general, one may expect that if there exists an edge (C_i, C_j) in G_{pag}, the interest in C_i can affect the interest in C_j (and vice-versa) proportionally to the weight of edge (C_i, C_j), although this is not always the case. Therefore, this analysis can be enhanced by looking at additional data, such as: (a) The net interest for each concept with respect to the other; (b) the presence of the two concepts in each shared page; (c) the distribution of the users on the pages containing C_i and/or C_j.
>
> **Task 2: Discovering relevant correlations between concepts.** If in G_{usr} there is an edge (v_i, v_j) with a high gross weight, we can expect that many users that look for concept C_i are also interested in C_j. This data may indicate a strong correlation between the two concepts from the user interest point of view. However, if C_i and C_j share some pages, this conclusion may be not reliable. To strengthen the analysis in this case, one can consider additional data, such as: (a) The net weight of edge (v_i, v_j) in G_{usr}; (b) the presence of the two concepts in each shared page; (c) the distribution of the users on the pages containing C_i and/or C_j.

In the next subsection we define visualization and interaction paradigms for graphs G_{pag}, G_{usr}, and G_{vis}. These paradigms are conceived to support the analyst in discovering relevant concepts (Task 1) and relevant correlation between concepts (Task 2).

3.1 Visualization and Interaction Paradigms

For a given day d, we display graphs G_{pag} and G_{usr} using the following visual paradigm (see Figure 1(a)):

- We compute two different drawings D_{pag} and D_{usr} for G_{pag} and G_{usr}, respectively. To help in the visual comparison of these drawings, corresponding vertices receive the same coordinates in D_{pag} and in D_{usr}, up to a translation of one of the two drawings. Also, each vertex v_i receives a label that describes the concept C_i.
- In each drawing, a vertex v_i is represented as a circle or as box, whose size is proportional to the gross interest of the corresponding concept C_i during day d, i.e., the size of v_i is proportional to $GrossInt(C_i, d)$.
- Edges in D_{pag} and in D_{usr} are colored differently. We assume that the edges of D_{pag} are colored blue and that the edges of D_{usr} are colored red. Each edge receives a label that shows its weight and/or it has a thickness proportional to its weight. In drawing D_{usr}, the weight initially used for an edge is its gross weight.

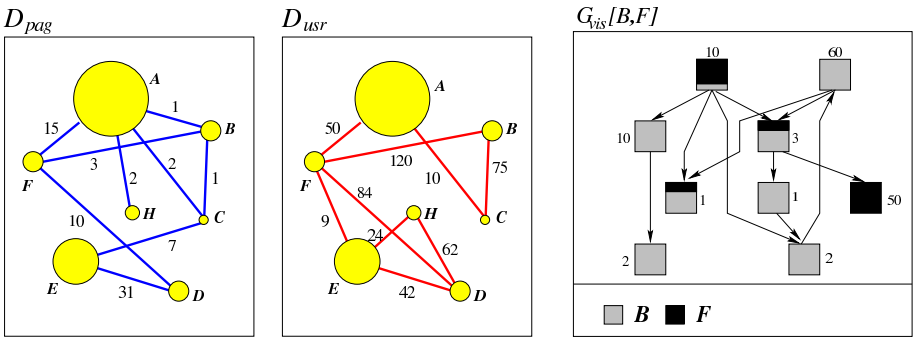


Figure 1: (a) Schematic illustration of the visual paradigm for G_{pag} and G_{usr}. (b) Schematic illustration of the visual paradigm for $G_{vis}[C_i, C_j]$.

The analyst can interact with the visualization by means of three main primitives:

Edge Hiding: This primitive allows the analyst to temporarily hide an edge in D_{usr} or in D_{pag} (or to restore a previously hidden edge). Hiding an edge from D_{usr} has no other effect in the visualization; this primitive can be useful to simplify the amount of data currently displayed, by hiding for example those edges having low weight. Hiding an edge (v_i, v_j) from D_{pag} can be used to get data according to point (a) of Tasks 1 and 2, as it has two effects in the visualization: (i) The size of vertex v_i becomes proportional to the net interest of C_i with respect to C_j, and the size of vertex v_j becomes proportional to the net interest of C_j with respect to C_i; (ii) the edge (v_i, v_j) in D_{usr} (if such an edge existed) is updated with its net weight.

Redraw: With this primitive the analyst can ask for an update of the current visualization for G_{pag} and G_{usr}, after some edges have been hidden (or restored)

from the last computation. Drawings D_{usr} and D_{pag} are recomputed with two main goals in mind: (i) improving the readability of the current visualization; (ii) preserving the user's mental map.

Edge Exploration: This primitive makes it possible to get data according to points (b) and (c) of Tasks 1 and 2, with respect to a selected edge (v_i, v_j) (either in G_{pag} or in G_{usr}). These data are displayed using a suitable visualization for G_{vis}. Namely, denote by $G_{vis}[C_i, C_j]$ the subgraph of G_{vis} induced by the nodes u_k such that page p_k contains C_i and/or C_j. We display $G_{vis}[C_i, C_j]$ as follows (see Figure 1(b)): (i) Each node u_k is represented as a box filled with two colors, one for C_i and the other for C_j, such that the amount of each color reflects the relative presence of the corresponding concept in page p_k. (ii) Each node u_k is associated with an integer label, showing how many users visited p_k.

4 COWA: COnceptual Web site Traffic Analyzer

Here we describe our software system, called *COWA*, for the visual analysis of concept-based Web site traffic data. The implementation is in Java, and we describe its main functionalities in Subsection 4.1. Subsection 4.2 describes the graph drawing algorithms of COWA.

4.1 Functionalities and Interaction Methods

COWA allows the analyst to define a set of concepts she is interested in, and a time interval T for which visitors' requests must be analyzed. Each concept C_i is manually defined by the analysts by inserting a profile, i.e., a brief textual description for C_i and a collection of terms characterizing C_i. After this phase, the system analyzes the Web server log file, and computes and stores the presence, consultation, and gross interest for each concept as defined in Subsection 2.1, for each day of T. It also computes graphs G_{pag} and G_{usr} for each day of T. Graph G_{vis} is computed on-line according to the analyst's requests.

Once all data needed for the analysis have been processed, the system shows a graphical interface, in which it is possible to focus on any day d in the time interval T. For a selected day d, the system visualizes the drawings D_{pag} and D_{usr}, according to the visual paradigm described in Section 3. In our implementation edges are represented with a discrete set of thickness values based on their weights. COWA integrates all the interaction primitives described in Subsection 3.1 and some additional functionalities, which are described below:

Node info: When passing the mouse over a node v_i of D_{pag} or of D_{usr}, the system displays a tool-tip that lists details, like the currently shown interest for C_i (i.e., its gross interest without the contribution of shared pages represented by some incident edges that have been hidden, if any), the total number of users who visited pages containing C_i and the total number of pages containing C_i, during day d.

Edge info: When passing the mouse over an edge (v_i, v_j) of D_{usr}, the system displays a tool-tip that reports its current weight (gross or net).

Drawing translation: The analyst can translate each drawing independently, to decide on the preferred relative positions of the given drawings for an effective visual comparison.

4.2 Graph Drawing Algorithms

Our drawing algorithms have to respect the visual paradigm described in Section 3. For a pair of selected concepts C_i, C_j, graph $G_{vis}[C_i, C_j]$ is represented as a layered drawing with a variant of the popular Sugiyama algorithm, described by Buchheim, Jünger and Leipert [4]; it uses only two bends per edge, which improves the readability of the drawing in most cases.

The most challenging task is computing a drawing of $G_{pag} = (V, E_{pag})$ and $G_{usr} = (V, E_{usr})$ in such a way that shared substructures can be easily identified by a visual inspection. Once the two drawings are computed, one of them can be translated using a split view visualization scheme (see, e.g., [16]).

A *simultaneous embedding* of $\{G_{pag}, G_{usr}\}$ is a drawing D_{pag} of G_{pag} and a drawing D_{usr} of G_{usr} where each vertex $v \in V$ is drawn at the same location in D_{pag} and in D_{usr}. The simultaneous embedding problem of a pair (or sequence) of graphs has a rich tradition in the graph drawing literature, especially when the graphs to be drawn are planar (see, e.g., [3, 12, 14, 15, 16]). For non-planar graphs, force-directed techniques are used by Erten *et al.* to define visualization schemes and design drawing heuristics where vertices shared by many graphs of the sequence are located in the center of the layout and edges shared by many graphs are not too long [16]. Kobourov and Pitta describe an interactive multi-user environment for simultaneous embedding where the number of edge crossings is kept under control by means of heuristic primitives [23]. Chimani *et al.* further investigate the problem of minimizing the number of crossings in a simultaneous embedding; they show the NP-hardness of the question and experimentally compare heuristics and exact drawing algorithms [7].

Similar to previous approaches, we use force-directed techniques to compute a simultaneous embedding of G_{pag} and G_{usr}. However, since G_{pag} and G_{usr} share the whole vertex set, the centrality of the shared vertices (considered in [16]) is not a meaningful aesthetic requirement in our case. Also, motivated by recent cognitive experiments, we do not insist on minimizing the number of edge crossings (as done in [7, 23]) but rather we try to control the visual quality of the edge crossings. More precisely, human-computer interaction experiments have shown that orthogonal crossings do not inhibit human task performance when reading a drawing and that users have a geodesic tendency when discovering relations between pairs of vertices [20, 21, 22]. Therefore, we compute a simultaneous embedding of $\{G_{pag}, G_{usr}\}$ by taking into account the following aesthetic requirements:

- The *crossing resolution* of any two edges of E_{usr} (E_{pag}) should be maximized. The crossing resolution is the minimum angle at which any two edges cross. Note that, to achieve a good crossing resolution some of the edges may be required to bend [13].

- The *geodesic edge tendency* should be guaranteed. Namely, for each edge $e = (u, v)$ with bends, we wish that the maximum distance from a point of e and the straight-line segment $\overline{uv}$ is minimized, and that e is monotone in the direction of $\overline{uv}$.

The drawing technique that we are going to describe further helps the user in the visual analysis of D_{pag} and D_{usr} by guaranteeing that edges of $E_{usr} \cap E_{pag}$ have the same representation in both drawings and by making the distance of adjacent vertices inversely proportional to the weights of their connecting edge, i.e., to the strength of their relationship.

Force-Directed Drawing Algorithm

We consider the graph $G = G_{pag} \cup G_{usr}$, compute a drawing D of G, and construct the wanted simultaneous embedding by extracting D_{pag} and D_{usr} from D. The edges of G are assigned one or two colors in the set $\{red, blue\}$. An edge e of G is colored blue if $e \in E_{pag} - E_{usr}$ and red if $e \in E_{usr} - E_{pag}$. Edges that belong to both G_{pag} and G_{usr} are blue and red at the same time. Our drawing algorithm works in three steps.

Step 1: An initial drawing of G is computed by applying a classical *spring-embedding* algorithm, where vertices are modeled as charged particles and edges are modeled as springs [11]. In our case, the charge of each vertex is proportional to the weight of the vertex, and the zero-energy length of each spring is inversely proportional to the weight of the corresponding edge. The weight of an edge e of G depends on its color: If e is blue its weight is the one that it has in G_{pag}; if it is red its weight is the one that it has in G_{usr}; if e is both red and blue, its weight is the maximum of its corresponding weights in G_{pag} and G_{usr}.

Step 2: Let D_1 be the drawing of G at the end of Step 1. Notice that, D_1 is such that vertices connected by edges with larger weights tend, if possible, to be closer to each other than vertices connected by edges with smaller weights. The graph is augmented by splitting some of its edges and by adding some new edges. Let $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ be a pair edges of D_1 such that e_1 and e_2 are both in G_{pag} or both in G_{usr} and such that e_1 and e_2 cross at a point c. Define a disc δ centered at c such that δ does not intersect in D_1 any edge distinct from e_1 and e_2, and having the same color as e_1 and e_2. For each value $i \in \{1, 2\}$, let p_i and q_i be the intersection points between δ and e_i (see Figure 2). Edge e_i is split into the path $(u_i, p_i), (p_i, q_i), (q_i, v_i)$. All edges of this path are given the same color as e_1 and e_2. Also, the straight-line edges $(p_1, p_2), (p_2, q_1), (q_1, q_2), (q_2, p_1)$ are added to the drawing; they are also given the same color as e_1. The four-cycle formed by these dummy edges is called the *cage* of crossing c. This procedure is iteratively applied to all crossings of the drawing.

Step 3: Let D_2 be the augmented drawing at the end of Step 2 and let G_2 be the corresponding augmented graph. The force-directed algorithm is executed again on G_2. The initial position of the vertices is the same as in D_2. The zero-energy length of the edges of all cages is fixed and made much smaller than the

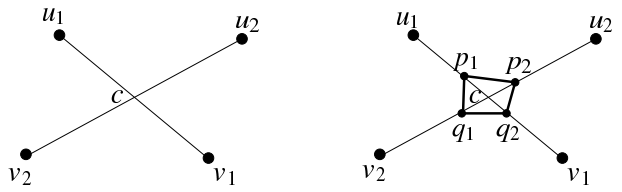


Figure 2: Illustration of Step 2 of the drawing algorithm.

zero-energy length of the shortest edge of D_1. Intuitively, this choice guarantees that the edge with bends will not be much longer than the straight-line segment connecting its endpoints and therefore ensures a good geodesic edge tendency. The spring stiffness of the edges of the cages is much larger than the stiffness of all other edges. This choice enforces the cages to be drawn as close as possible to squares, which guarantees that their diagonals make a crossing approximating right angles (i.e., the portions of the edges of G inside the cage intersect each other orthogonally). At the end of this step the dummy edges are removed, the dummy vertices are replaced with bends, and the edges are smoothed and drawn as Bézier curves (bends are used to define the control points).

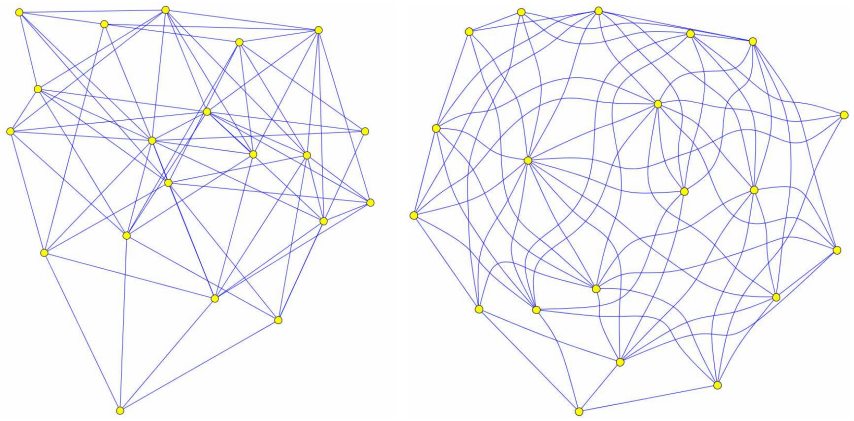
4.3 Experimental Analysis

We implemented our modified spring embedding algorithm and ran an experimental evaluation that compares it with a classical spring embedding algorithm in terms of crossing resolution and geodesic edge tendency, independently of the fact that we also apply this heuristic to simultaneously embed two graphs. Note that, a direct comparison with other simultaneous embedding heuristics that minimize crossings is not meaningful, since our goal is the optimization of different objective functions.

Figure 3(a) shows an example of a drawing of a graph with 20 vertices and 72 edges at the end of Step 1, that is, once a classical spring embedding algorithm has been applied to draw it. Figure 3(b) is a drawing of the same graph once Steps 2 and 3 have been applied. A similar comparison is shown in Figure 4(a) and Figure 4(b) for a graph of 50 vertices and 75 edges.

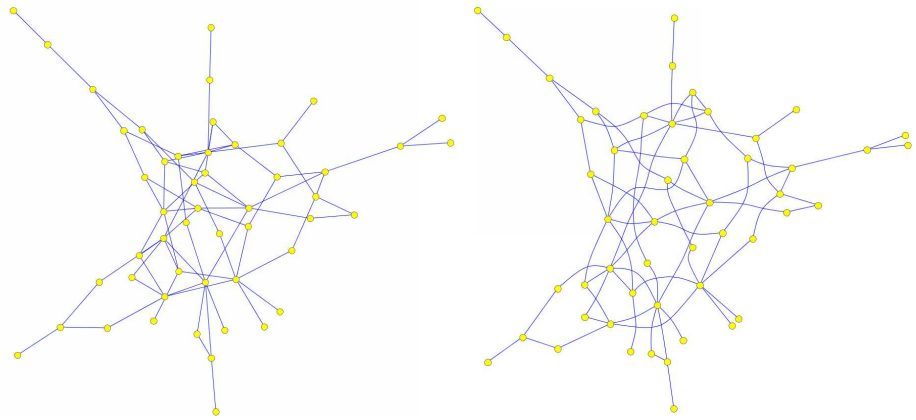
The experiments were executed on a set of 300 graphs with up to 50 vertices (in the context of our application, it is reasonable to assume that an analyst cannot effectively deal with too many concepts at the same time). For each number $n \in \{10, 20, \ldots, 50\}$, we randomly generated 60 graphs with n vertices, and a density value d randomly chosen in the range $[2 - 4]$ for graphs up to 20 vertices, and in the range $[1.5 - 2.5]$ for graphs with more than 20 vertices. For each pair $\langle n, d \rangle$, a graph with n vertices and $m = d \cdot n$ edges was generated with a uniform probability distribution. We now discuss the experimental results.

Running Time. The theoretical time complexity of our technique increases with respect to a standard force-directed algorithm, due to the potentially high number of dummy vertices introduced in Step 2. However, we observed in



(a) (b)

Figure 3: Drawing of a graph with 20 vertices and 44 edges at the end of: (a) Step 1; (b) Step 3.

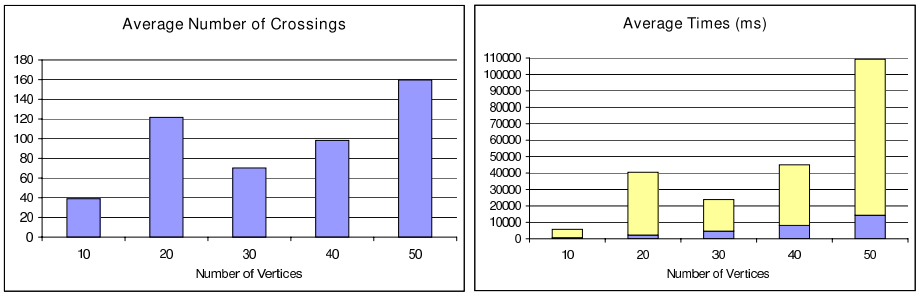


(a) (b)

Figure 4: Drawing of a graph with 50 vertices and 75 edges at the end of: (a) Step 1 (b) Step 3.

practice that a small number of iterations in Step 3 is typically sufficient to get highly readable drawings (as described later), and then the running time of the whole algorithm remains feasible for relatively small graphs (like those arising in our application). Figure 5(a) shows the average number of crossings of the drawings computed at the end of Step 1. Figure 5(b) shows the average running time in milliseconds required by the algorithm, by showing separately the time taken by Step 1 (dark bars) and by Steps 2 and 3 (light bars). In our

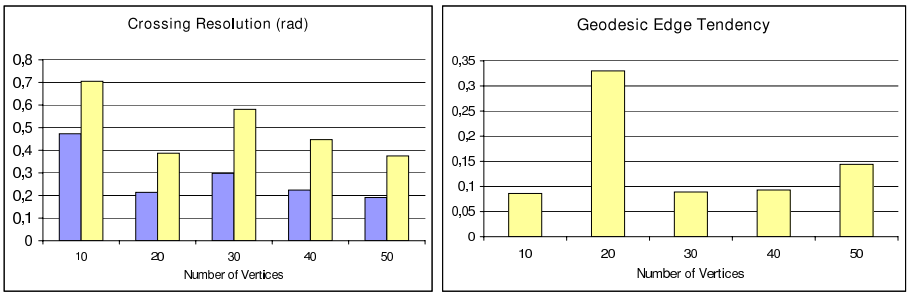
implementation, the force-directed algorithm executes in Step 3 only 10% of the number of iterations executed in Step 1.



(a) (b)

Figure 5: (a) Number of crossings at the end of Step 1. (b) Average running time of the algorithm. The dark bars indicate the time required by Step 1, the light bars the one required by Steps 2 and 3.

Crossing Resolution. As mentioned before, the crossing resolution is defined as the minimum angle at which any two edges cross. Figure 6(a) reports the crossing resolution (averaged over the instances with the same number of vertices) for the drawings obtained at the end of Step 1 and at the end of Step 3. From the chart it is possible to observe that Step 3 dramatically improves the crossing resolution of the drawings. The improvement is between 50% and 100%; the average crossing resolution of the final drawings is always greater than 30°.



(a) (b)

Figure 6: (a) Crossing resolution at the end of Step 1 (dark bars) and at the end of Step 3 (light bars). The angles are in radians. (b) Geodesic edge tendency.

Geodesic Edge Tendency. In order to evaluate the geodesic edge tendency of a drawing, for each edge (u, v) with bends, we measured the ratio between the maximum distance of (u, v) from the straight-line segment $\overline{uv}$ and the length of $\overline{uv}$; we took the average value over all edges of the drawing. On average, the

maximum distance of an edge (u, v) from segment $\overline{uv}$ is rather small (less than 12% of the length of $\overline{uv}$), which implies a good geodesic edge tendency for the bent edges. We also measured the monotonicity of the edges according to the direction of $\overline{uv}$, and we found that only 4 edges in all computed drawings are not monotone (this number represents a percentage very close to zero if compared with the total number of edges in all drawings).

Finally, we observe that introducing bends along the edges has additional advantages in terms of drawing readability. Namely, Finkel and Tamassia have shown that using bends along the edges in a force-directed method usually improves the edge separation and the angular resolution of the layout [18] (see, e.g., Figures 3 and 4). The main difference between our approach and that of Finkel and Tamassia is that their technique introduces one or two bends along each edge before executing the force-directed algorithm and does not have control on the crossing resolution; our approach places bends close to edge crossings and controls crossing resolution and geodesic edge tendency.

5 Case Studies

We applied COWA to the analysis of visitor's requests for two popular Web sites, the first containing information about volleyball championships and teams and the second containing national and international news [2].

5.1 Interest Analysis of a Volleyball Web Site

Types of analysis like the one described in the following may greatly help volleyball managers to define better focused communication strategies and/or a differentiation of investments to improve the popularity of the sport.

The Web server log file of the analyzed site contains about 4,000,000 of entries each day, which correspond to several thousands of visitors (a single user access typically causes several HTTP requests automatically executed by the user's browser to load the different parts of the final content displayed to the user). We focused on the interval time T = [1 April 2009 - 13 April 2009], and analyzed the following ten concepts: *Series A, Series B, National Cups, International Cups, Senior Male National Team, Senior Female National Team, Junior Male National Team, Junior Female National Team, Beach Volley, Schools.* After a preliminary analysis of the page contents, we decided to describe each concept by using a profile consisting of two or three terms.

Table 1 reports the gross interest of each concept in each day of the period. In general, we can observe that for most of the concepts there is a homogeneous trend, and that concept *International Cups* is the one with the largest gross interest. However, there are days in which the gross interest increases significantly for some specific concepts. For example on April 12, we have a relevant increase of gross interest for the concepts *Series A* with respect to its average

[2] We do not report the URLs and the name of these Web sites for reasons of privacy.

Concept / Day	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Series A	4	4.3	3.3	2.9	5.3	4.9	3.5
Series B	1.8	2	1.9	4.6	6.6	4.1	2.2
National Cups	3.6	4.2	2.8	2.4	3.8	3.6	4.5
International Cups	68.3	70.4	67.9	72.5	78.8	87.5	86.4
Senior Male National Team	10.5	9.6	8.1	6.2	6.2	7.9	10
Senior Female National Team	35.3	35.4	33.8	33.6	32.6	39.9	44.7
Beach Volley	7.9	7.3	6	2.6	1.6	3.3	4.9
Junior Male National Team	9.5	8.6	7.2	5.2	5.3	7	8.9
Junior Female National Team	8.9	8.8	7.3	5.4	5.4	7.3	9.2
Schools	29.6	27.5	23.7	13.9	11.4	22.2	28

Concept / Day	**8**	**9**	**10**	**11**	**12**	**13**
Series A	3.5	4.8	6.5	8	11.3	8.5
Series B	2	2.4	2	2.6	2.9	2.9
National Cups	6.7	11	7.1	11.4	12.3	8.4
International Cups	82.6	75.7	80.5	79.7	84.2	77.7
Senior Male National Team	9.7	11.7	14.4	9.7	12.3	12
Senior Female National Team	42.5	38.6	40.2	37.4	35.4	35.6
Beach Volley	4.8	7.2	6.4	6.4	5	7.4
Junior Male National Team	9.2	11.2	13	9.2	11.9	11
Junior Female National Team	9	10.3	12.3	9	10	11
Schools	25.7	26.6	25.9	19.6	17.4	25

Table 1: Gross interests of the volleyball Web site concepts in the interval time [1 April 2009 - 13 April 2009].

interest in T. Checking the volleyball events on a calendar, we discovered that the matches of the second division of Series A took place on April 11. This probably means that many users accessed the site the day after to check the results and the new ranking of the teams.

We asked the system to visualize graphs G_{pag} and G_{usr} in a specific day, namely April 1. The drawings D_{pag} and D_{usr} of these two graphs computed by using our drawing algorithm are shown in Figure 7(a). The sizes of the nodes visually convey the information on the gross interest of each concept much more rapidly and intuitively than looking at a table. Also, we can visually compare the two drawings and interact with them in order to mine additional information that cannot be easily extracted with the concept-based audience metric model by Norguet *et al.* [26, 27].

Concerning Task 1 of Section 3 (discovering relevant concepts) and looking at D_{pag}, one can study to what extent the interest in a certain concept is affected by some other (i.e., how much the gross interest of a certain concept is close to its net interest with respect to some other concepts). For example,

hiding edge (*Series A, International Cup*), we obtain two drawings showing the relative net interests of the two concepts, and the net interest of edge (*Series A, International Cup*) in D_{usr} (see Figure 7(b)). We can see that concept *International Cup* is affected by concept *Series A* in a remarkable way; indeed, most of the users that visited these two concepts on April 1 always visited at least one page in which the two concepts appear simultaneously. This implies that if we want to know what is the real interest of the visitors for the *International Cups* independent of their interest for the *series A*, we have to look to the size of this concept in Figure 7(b) more than in Figure 7(a).

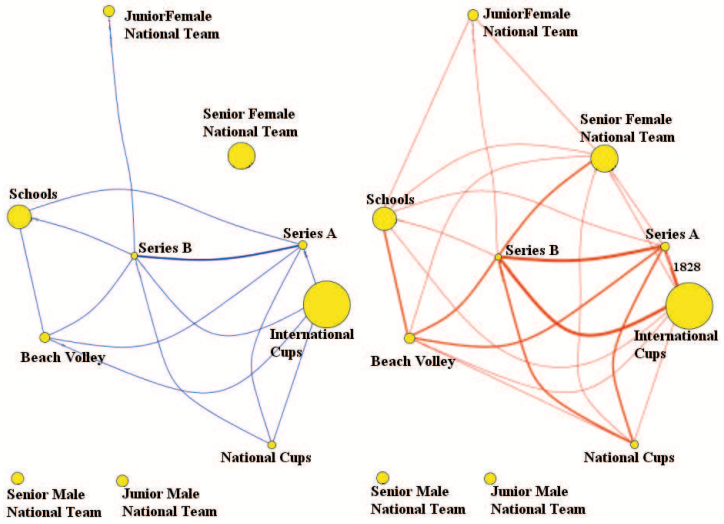
More details on the distribution of the visitors on the two concepts can also be obtained by looking at the drawing of G_{vis}[*Series A*, *International Cup*] (see Figure 8(a)). Recall that each node in G_{vis} represents a page containing one or both the two concepts. A label is shown for each node, indicating the number of distinct accesses. The dark color in each page indicates the relative relevance of concept *Series A*, the light color indicates the relative relevance of concept *International Cup*. From the picture, we can observe that there are few pages that share the two concepts, but most of the users visited them during their browsing. This data strengthens the fact that the interest in one concept cannot be considered independently of the other.

Concerning Task 2 of Section 3 (discovering relevant correlations between concepts), there are several edges in D_{usr} that connect concepts not adjacent in D_{pag}, and some of these edges have a high weight. For example, many users that visited concept *Senior Female National Team* were also interested in *Series B* and *Series A*. This indicates a reliable connection between such pairs of concepts in the considered day; volleyball managers can therefore deduce that during this day the visitors were mainly interested in the series A and B of the senior female teams. We also note that there were several users interested in both *Series A* and *School*, which are however adjacent concepts in D_{pag}. Hiding edge (*Series A*, *School*) in D_{pag} we can get the net interests for the edge and the relative net interests for the two concepts (see Figure 8(b)). In this case it is worth observing that they are rather heavily affecting each other and 9 users out of 14 users visited only pages in which the two concepts do not appear simultaneously. Hence, we cannot completely trust the logical correlation between these two concepts.

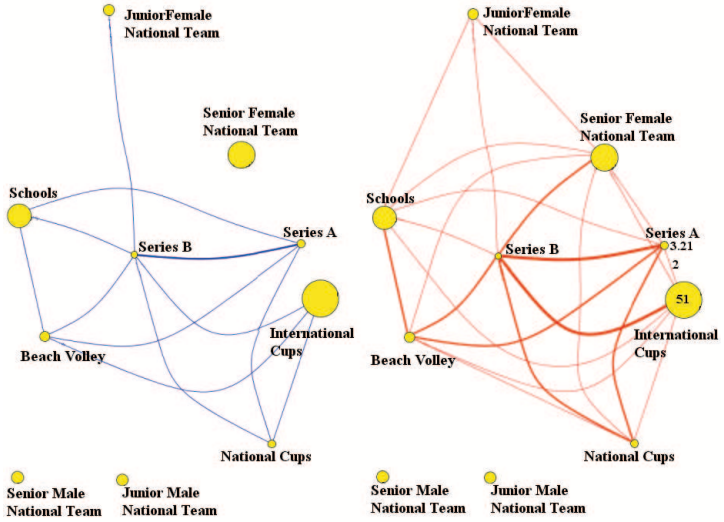
5.2 Interest Analysis of a News Web Site

For this site we considered the data of one day, June 17, 2009. In this day the site received about 11,000 accesses, corresponding to 4,500 users and 6,000 documents. We analyzed the following concepts: *Economy*, *Sport*, *Politics*, *Foreign Countries*, *Italy*, *Bari's Sex Scandal*, *Movies*, *Crime*. Again, each concept was described by a profile of three terms.

The drawings of the graphs D_{pag} and D_{usr} are shown in Figure 9(a). We can observe, from the gross weight of the nodes, that *Movies* and *Bari's Sex*



(a)

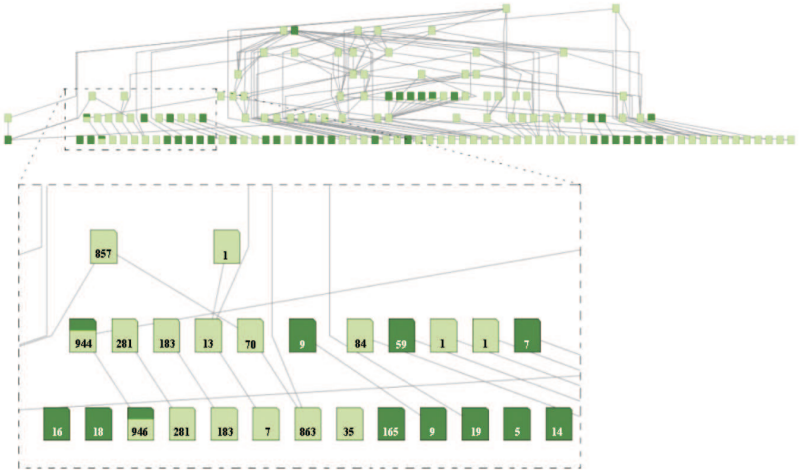


(b)

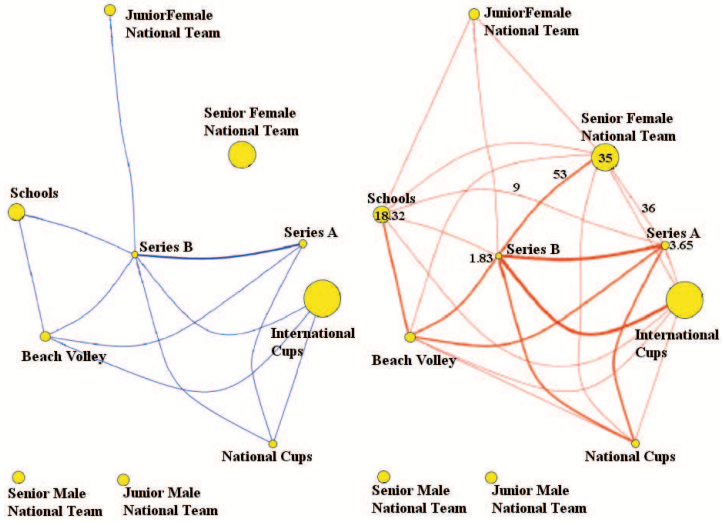
Figure 7: (a) Drawings D_{pag} (to the left) and D_{usr} (to the right) on April 1 for the volleyball web site. (b) The effect of hiding edge (*Series A, International Cups*).

Scandal have the highest weights[3]. As for the previous case study, we discuss several issues related to Tasks 1 and 2.

[3]Concept *Bari's Sex Scandal* was actually one of the news of major relevance in that period in the Italian newspapers

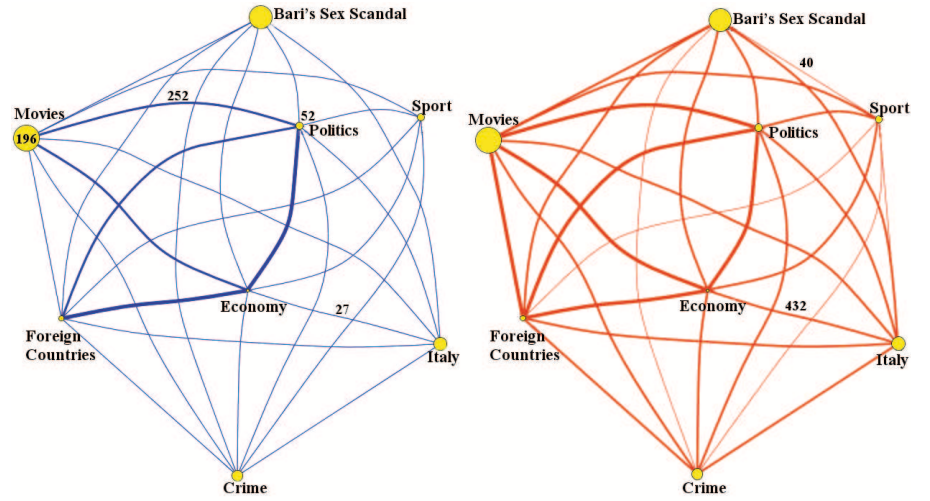


(a)

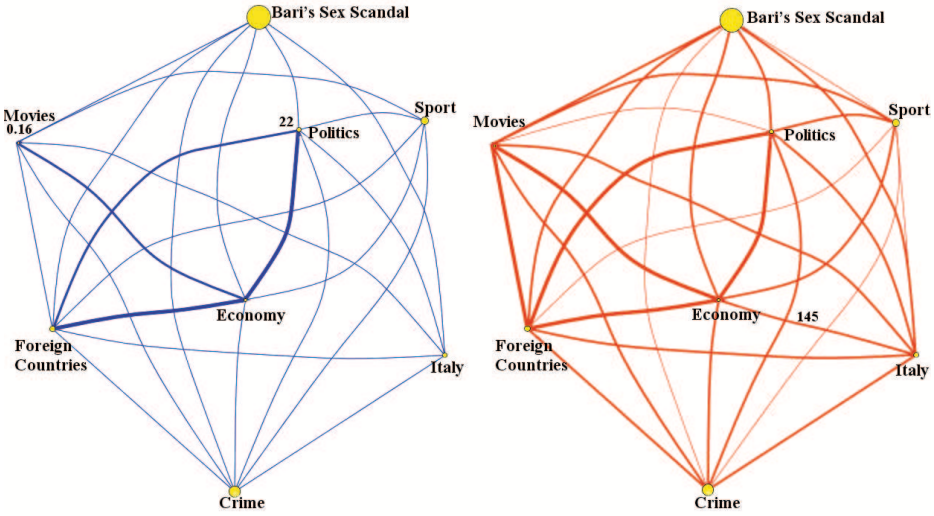


(b)

Figure 8: (a) Exploring graph $G_{vis}[(\textit{Series A, International Cups})]$. It is possible to see that the few pages in which the two concepts appear simultaneously are highly accessed. The dark color represents concept *Series A*, while the light color represents concept *International Cup*. (b) The effect of hiding edge (*Series A*, *School*). The logical connection between the two concepts *Series A* and *School* is not reliable.



(a)



(b)

Figure 9: (a) Drawings D_{pag} (to the left) and D_{usr} (to the right) for the news web site. (b) The effect of hiding edge (*Politics, Movies*) and edge (*Economy, Italy*).

Concerning Task 1 and looking at D_{pag}, we note for example that the edge between concepts *Politics* and *Movies* has a weight of 252, which implies a high number of shared pages between these two concepts in the site. The gross interests for the two concepts are 52 and 196, respectively. If we hide edge (*Politics*,

Movies) (see Figure 9(b)) the net interest of *Movies* becomes negligible, that is, less than 0.2, which means that the gross interest for this concept is mainly caused by the accesses to the pages shared with concept *Politics*. However, the big difference between the gross interests of *Politics* and *Movies* also implies that in the pages shared by these concepts, *Movies* is predominant with respect to *Politics*, and therefore the interest of the users for concept *Movies* can be considered reliable. To further support this conclusion, we verified that in the first days of June 2009 on the newspapers it was announced the imminent release of a sequel for the popular movie *Transformers*.

Concerning Task 2, we identify pairs of correlated concepts. We notice that edge (*Sport*, *Bari's Sex Scandal*) appears in D_{usr} with weight 40, but it does not appear in D_{pag} (see Figure 9(a)). This immediately suggests that in June 17 many of the users that read about sports also read about Bari's sex scandal intentionally. Again, the edge (*Economy*, *Italy*) appears both in D_{pag} (with weight 27) and in D_{usr} (with weight 432). We verified that the total number of pages containing the concept *Economy* was 31, hence 87% of these pages intersect with pages containing concept *Italy*. In spite of this relevant intersection, hiding the edge (*Economy*, *Italy*) in D_{pag}, the net weight in D_{usr} is still relatively high, namely 145, i.e. 33.6% of the users that accessed both concepts (see Figure 9(b)). This suggests a real correlation between concepts *Economy* and *Italy* from a user interest perspective.

6 Conclusions and Open Problems

This paper presents a graph drawing application to Web site traffic analysis. Our contribution is twofold. From the point of view of data and knowledge engineering, we introduced new notions and models to measure the user interest into the concepts of a Web site. From the point of view of graph visualization, it presented new visual analytics techniques and an adaptation of force-directed heuristics to compute drawings of graphs where the angles formed by crossing edges are large and each bent edge is a monotone curve relatively close to the geodesic path between its end-vertices. These concepts were implemented and experimented within a system called COWA. The interaction with COWA makes it possible to measure the interest in a pair of concepts; it would be interesting to extend this facility and measure the interest for groups or clusters of concepts. The graph drawing heuristic draws edges having a bend for every crossing; it would be interesting to reduce the curve complexity of the edges by integrating our heuristic with crossing reduction techniques (see, e.g., [7]).

Acknowledgements

We acknowledge the anonymous reviewers for their valuable comments.

References

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[2] P. Berkhin, J. D. Becher, and D. J. Randall. Interactive path analysis of web site traffic. In *ACM SIGKDD*, pages 414–419, 2001.

[3] P. Braß, E. Cenek, C. A. Duncan, A. Efrat, C. Erten, D. Ismailescu, S. G. Kobourov, A. Lubiw, and J. S. B. Mitchell. On simultaneous planar graph embeddings. *Comput. Geom.: Theory and Appl.*, 36(2):117–130, 2007.

[4] C. Buchheim, M. Jünger, and S. Leipert. A fast layout algorithm for k-level graphs. In *GD 2000*, volume 1984 of *LNCS*, pages 229–240, 2000.

[5] J. Chen, L. Sun, O. Zaïane, and R. Goebel. Visualizing and discovering web navigational patterns. In *7th International Workshop on the Web and Databases (WebDB 2004)*, pages 13–18, 2004.

[6] J. Chen, T. Zheng, W. Thorne, O. R. Zaïane, and R. Goebel. Visual data mining of web navigational data. In *11th International Conference on Information Visualization (IV)*, page 649656, 2007.

[7] M. Chimani, M. Jünger, and M. Schulz. Crossing minimization meets simultaneous drawing. In *IEEE PacificVis 2008*, pages 33–40, 2008.

[8] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24(1):305339, 2005.

[9] P. Cimiano, S. Staab, and J. Tane. Deriving concept hierarchies from text by smooth formal concept analysis. In *"GI Workshop Lehren Lernen - Wissen - Adaptivität*, pages 72–79, 2003.

[10] R. Cooley. The use of web structure and content to identify subjectively interesting web usage patterns. *ACM Trans. Internet Techn.*, 3(2):93–116, 2003.

[11] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph Drawing*. Prentice Hall, Upper Saddle River, NJ, 1999.

[12] E. Di Giacomo and G. Liotta. Simultaneous embedding of outerplanar graphs, paths, and cycles. *Intern. Journ. of Comput. Geom. and Appl.*, 17(2):139–160, 2007.

[13] W. Didimo, P. Eades, and G. Liotta. Drawing graphs with right angle crossings. In *Algorithms and Data Structures Symposium*, volume 5664 of *LNCS*, pages 206–217, 2009.

[14] C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. V. Yee. Graphael: Graph animations with evolving layouts. In *GD 2003*, volume 2912 of *LNCS*, pages 98–110, 2003.

[15] C. Erten and S. G. Kobourov. Simultaneous embedding of a planar graph and its dual on the grid. *Theory Comput. Syst.*, 38(3):313–327, 2005.

[16] C. Erten, S. G. Kobourov, V. Le, and A. Navabi. Simultaneous graph drawing: Layout algorithms and visualization schemes. *J. Graph Algorithms Appl.*, 9(1):165–182, 2005.

[17] D. Fensel. *A Silver Bullet for Knowledge Management and Electronic Commerce*. Springer Verlag, 2000.

[18] B. Finkel and R. Tamassia. Curvilinear graph drawing using the force-directed method. In *GD 2004*, volume 3383 of *LNCS*, pages 448–453, 2004.

[19] B. Ganter and R. Wille. *Formal Concept Analysis Mathematical Foundations*. Springer Verlag, 1999.

[20] W. Huang. Using eye tracking to investigate graph layout effects. In *Asia-Pacific Symposium on Visualization*, pages 97–100, 2007.

[21] W. Huang, P. Eades, and S.-H. Hong. Graph reading behavior: Geodesic-path tendency. In *IEEE PacificVis 2009*, pages 137–144, 2009.

[22] W. Huang, S.-H. Hong, and P. Eades. Effects of crossing angles. In *IEEE PacificVis 2008*, pages 41–46, 2008.

[23] S. G. Kobourov and C. Pitta. An interactive multi-user system for simultaneous graph drawing. In *GD 2004*, volume 3383 of *LNCS*, pages 492–501, 2004.

[24] P. Kolari and A. Joshi. Web mining: Research and practice. *Computing in Science and Engg.*, 6(4):49–53, 2004.

[25] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[26] J.-P. Norguet, B. Tshibasu-Kabeya, G. Bontempi, and E. Zimanyi. A page-classification approach to web usage semantic analysis. *Engineering Letters*, 14(1), 2007.

[27] J.-P. Norguet, E. Zimányi, and R. Steinberger. Improving web sites with web usage mining, web content mining, and semantic analysis. In *SOFSEM 2006*, volume 3831 of *LNCS*, pages 430–439, 2006.

[28] Z. Pabarskaite and A. Raudys. A process of knowledge discovery from web log data: Systematization and critical review. *J. Intell. Inf. Syst.*, 28(1):79–104, 2007.

[29] V. Pascual-Cid. An information visualisation system for the understanding of web data. In *IEEE Symp. on Visual Analytics Science and Technology (VAST'08)*, pages 183–184, 2008.

[30] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations*, 1(2):12–23, 2000.

[31] A. Youssefi, D. Duke, and M. Zaki. Visual web mining. In *13th international World Wide Web conference on Alternate track papers & posters (WWW'04)*, pages 394–395, 2004.