Volume 46, issue 3, May 2010

ISSN 0306-4573

INFORMATION PROCESSING & MANAGEMENT

An International Journal

Email: ipm@unisi.ch
Web site: http://www.elsevier.com/locate/infoproman

Information Processing and Management 46 (2010) 284–294

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Personalization by website transformation: Theory and practice

Saverio Perugini [*]

Department of Computer Science, University of Dayton 300 College Park, Dayton, OH 45469–2160, USA

ARTICLE INFO

Article history:
Received 6 October 2008
Received in revised form 18 December 2009
Accepted 26 December 2009
Available online 12 February 2010

Keywords:
Hierarchical hypermedia
Information personalization
Navigation
Out-of-turn interaction
Website transformation

ABSTRACT

We present an analysis of a progressive series of out-of-turn transformations on a hierarchical website to personalize a user's interaction with the site. We formalize the transformation in graph-theoretic terms and describe a toolkit we built which enumerates all of the traversals enabled by every possible complete series of these transformations in any site and computes a variety of metrics while simulating each traversal therein to qualify the relationship between a site's structure and the cumulative effect of support for the transformation in a site. We employed this toolkit in two websites. The results indicate that the transformation enables users to experience a vast number of paths through a site not traversable through browsing and demonstrate that it supports traversals with multiple steps, where the semblance of a hierarchy is preserved, as well as shortcuts directly to the desired information.

1. Introduction

Personalization refers to automatically customizing interactive information systems based on user preferences. Personalization technologies are now widely utilized on the web. While most approaches to personalization are either template-based (i.e., slot fillers such as those found at *My Yahoo!* (Manber, Patel, & Robinson, 2000)) or artificial intelligence-oriented, the central theme of our approach is to personalize a user's interaction with a website by progressively transforming its structure in response to every user interaction in a session with the site to help the user experience paths through the site not traversable through browsing. For instance, consider a user shopping for a book by Aldous Huxley at a website which only presents books by genre. Such a user unsure in which genres Huxley published is forced to browse through all genres to manually find books of interest. While this user is unable to respond to the current solicitation for input (i.e., genre), she does have information (i.e., author) relevant to the information-seeking task even though that information is not required until the user is nested deeper into the catalog.

Our approach to this problem is a technique called *out-of-turn interaction*. The idea is to permit a user navigating a hierarchical website to postpone clicking on any of the hyperlinks presented on the current page (e.g., when unable or unwilling to respond to the current prompt for input) and, instead, communicate the label of a hyperlink nested deeper in the hierarchy. When the user supplies such out-of-turn input we transform the hierarchy to reflect the user's informational need. In the example above, when unsure in which genres Huxley published, the user may communicate

[*] Tel.: +1 9372294079; fax: +1 9372292193.
E-mail address: saverio@udayton.edu
URL: http://academic.udayton.edu/SaverioPerugini

0306-4573/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.12.008

'Aldous Huxley' to the site out-of-turn. In response, we would transform the hierarchical organization of the catalog so that all hyperlinks leading to books not written by Huxley are purged and re-present the hierarchy to the user. As a result of the transformation, the user would see a page of hyperlinks representing genres. However, each hyperlink remaining would eventually lead to a book by Huxley. Thus, out-of-turn interaction permits the user to circumvent any intended flows of navigation hardwired into the hyperlink structure by the designer and, in this manner, helps reconcile any mismatch between the site's one-size-fits-all organization and the user's model of information seeking.

We built a transformation engine as a web service based on this idea which prunes a hierarchical site when given out-of-turn input. We also built two interfaces to communicate the input to the engine: a voice interface, implemented with VoiceXML and X+V, which permits the user to supply out-of-turn inputs through speech and enables multimodal interaction when used in conjunction with hyperlinks, and *Extempore*, implemented with XUL, which is a cross-platform toolbar plugin embedded into the Mozilla Firefox web browser. The transformation engine, interfaces, and a coordinating interaction manager constitute a customizable software framework for creating web personalization systems with support for out-of-turn interaction (Narayan, Williams, Perugini, & Ramakrishnan, 2004). We have applied this technique to various websites, including the Open Directory Project, a large web directory.

We have studied out-of-turn interaction from software implementation (Narayan et al., 2004) and human-computer interaction (HCI) (Perugini, Anderson, & Moroney, 2007) perspectives. The goal of this paper is to study the transformation which supports this technique from a graph transformation perspective and analyze the traversals of the site it enables. This is an intermediate approach between the implementation and HCI complementary approaches. Specifically, we (i) formalize the transformation in graph-theoretic terms, (ii) describe a toolkit we built which computes and simulates all of the traversals enabled by all possible complete series of out-of-turn transformations in any site to qualify the relationship between how terms are distributed through the site's structure and the effect of support for the transformation in a site, and (iii) report the results of employing this toolkit in two websites.

The central mantra of this paper is that a series of website transformations on a site supports a set of traversals through the site we called an interaction paradigm:

$$\text{Transformation}(\cdots\ \text{Transformation}(\text{Website}, \text{Hyperlink label}), \cdots, \text{Hyperlink label})\ \Rightarrow\ \text{Interaction paradigm}.$$

Only a small subset of all possible traversals made possible by a series of out-of-turn transformations on a site can be experienced through browsing.

2. Related research

Traditionally, there are two main approaches to web personalization: template- and AI-oriented approaches. The template-based approach (Perugini & Ramakrishnan, 2003) (also called *checkbox* personalization) is predominately employed in the *my* sites (e.g., *My Yahoo!* (Manber et al., 2000) or *My eBay*). Most all e-commerce sites now provide such a facility. The onus is on the user to explicitly specify her preferences and, as a result, the content, structure, or presentation of the website is tailored accordingly. Such an approach involves explicit user modeling (Konstan et al., 1997). While template-based approaches to personalization do not suffer from privacy concerns, the level of personalization delivered is bounded by the investment of the user in communicating his interests, and often higher-order connections or serendipitous recommendations are not possible. On the other hand, AI-based approaches to web personalization involve covertly monitoring user behavior and activity, often through web usage mining (i.e., web log analysis) (Mobasher, Cooley, & Srivastava, 2000), to implicitly glean user preference and, ultimately, build a user model which is used as a basis from which to personalize the site. One popular example of such an approach is adaptive websites (Perkowitz & Etzioni, 2000). Unlike template-based personalization, the success of AI-oriented approaches is not predicated on the cooperation of the user. However, these methods are perceived as invasive and raise privacy concerns (Riedl, 2001). The primary enabling technology for these approaches is web mining (Eirinaki & Vazirgiannis, 2003; Kosala & Blockeel, 2000), and specifically web usage mining (Srivastava, Cooley, Deshpande, & Tan, 2000). This user-model through access monitoring approach is seen in the adaptive hypermedia (Brusilovsky, 2001) and interactive information retrieval (White, Jose, & Ruthven, 2006) communities.

The out-of-turn website transformation approach to personalized interaction does not fit into either of these categories. Rather, out-of-turn interaction can be broadly characterized as a faceted browsing and search technique (Hearst et al., 2002), and is particularly related to the *zoom* operation in *dynamic taxonomies* (Sacco, 2000). Faceted browsing and search (Sacco & Tzitzkas, 2009) seeks to marry navigational (e.g., *Yahoo!*) and direct (free form) search (e.g., *Google*), and has received an increased level of attention from the interactive information retrieval community recently as an approach between template- and AI-based techniques.

Faceted browsing and search permits a user to explore a multi-dimensional dataset in a manner which matches the user's mental model of information-seeking, thereby personalizing the user's interaction with site (e.g., 'You prefer to browse recipes using a by main ingredient, dish type, preparation method motif while I prefer to browse by dish type, preparation method, and main ingredient'). The multi-faceted index of recipes at http://epicurious.com is perhaps the most illustrative example of a faceted classification on the web (Hearst, 2000).

3. Theory: out-of-turn transformation formalism[1]

Fundamentally, the out-of-turn transformation is a closed transformation over a graph modeling the hyperlink structure of a website. In this section we discuss how websites can be represented as graphs, how interacting out-of-turn transforms a graph, and the implications a series of those transformations have on web interaction.

3.1. Websites as graphs

It is instructive to think of websites as graphs. For instance, Fig. 1 (left) illustrates a directed acyclic graph (DAG) model of a hierarchical website with characteristics similar to web directories such as the Open Directory Project (ODP) at http://dmoz.org. Edges help model paths through a website a user follows to access leaf vertices, which model leaf webpages containing content. We refer to a leaf content page as *terminal information* and the terms therein as units of terminal information. Edge-labels, which we refer to as *structural information*, model hyperlink labels or, in other words, choices made by a navigator en route to a leaf. An edge-label, a unit of structural information, is therefore a term of information-seeking (simply a *term* hereafter) which a user may bring to bear upon information seeking. Structural information thus helps make distinctions among terminal information.

A set of terms is *complete* when it determines a particular terminal webpage; otherwise it is *partial*. An *interaction set* of a DAG D is the complete set of the terms along a path from the root of D to a leaf vertex of D. An interaction set constitutes complete information; any proper subset of it is partial information. An interaction set of D classifies a leaf vertex of D, but does not capture any order of the terms therein. On the other hand, a *sequence* is a total order of an interaction set wrt the parenthood relation of the site. In other words, a sequence represents a path from the root to a leaf in a site. The sequence $\prec$shopping, apparel, winter$\succ$ is in the DAG shown in Fig. 1 (left). A term is *in-turn* information if it appears as a hyperlink label on the user's current webpage and is, thus, currently solicited by the system. On the other hand, a term is *out-of-turn* information if it represents a hyperlink label nested somewhere deeper in the site and is, thus, currently unsolicited from the system, but relevant to information-seeking. In any DAG, in-turn and out-of-turn information is mutually-exclusive.

3.2. Transformations

We now present some website transformations. *Term extraction* is a total function $TE : \mathscr{D} \rightarrow P(\mathscr{T})$ which given D returns the set of all unique terms in D, where $\mathscr{D}$ represents the universal set of DAGs, $\mathscr{T}$ represents the universal set of terms, and $P(\cdot)$ denotes the power set function. A *term-co-occurrence set* of D is a set $T \subseteq TE(D)$. Let the *level* of an edge-label in D be the depth of the source vertex of the edge it labels. If a given edge-label occurs multiple times in D, a level is associated with every occurrence. A *term-level set* of D then is a term-co-occurrence set comprising all unique terms in D with the same level. *Term-level extraction* is a total function $TLE : (\mathscr{D} \times \mathscr{N}) \rightarrow P(TE(D))$ which given D and a level $l(\geqslant 1) \in \mathscr{N} = \{1, 2, \ldots, M\}$ returns the set of all unique terms in D with level l (i.e., a term-level set), where M represents the maximum depth of D. If D represents the DAG in Fig. 1 (left), $TLE(D, 2) = \{\text{international},\ \text{advertising},\ \text{coupons},\ \text{electronics},\ \text{apparel}\}$.

In any DAG, $TLE(D, 1)$ returns the set of terms available to supply through browsing or, in other words, in-turn information. *Browse* is a partial function $B : (\mathscr{D} \times \mathscr{T}) \rightarrow \mathscr{D}_\perp$ which given D and a term $t \in TLE(D, 1)$ returns the sub-DAG rooted at the target vertex of the edge in D labeled with t whose source vertex is the root of D. If D is the DAG in Fig. 1 (left), $B(D, \text{shopping})$ returns the sub-DAG rooted at vertex 3, which represents the result of a user clicking on the hyperlink labeled 'shopping'. The symbol $\perp$ denotes the partial nature of the function (i.e., the value of B is undefined for some inputs). If $t \notin TLE(D, 1), B$ returns $\perp$.

Out-of-turn transformation is a partial function $OOT_1 : (\mathscr{D} \times \mathscr{T}) \rightarrow \mathscr{D}_\perp$ which given D and a term $t \in TE(D)$ returns D':

$$OOT_1(D,\ t) = CE(\overbrace{\underbrace{BP(D,\ \overbrace{FP(D,\ t)}^{\text{Fig. 1 (left)}})}_{\text{Fig. 1 (center)}},\ t\,)}^{\text{Fig. 1 (right)}}, \tag{1}$$

where

- FP (*forward propagate*): $(\mathscr{D} \times T) \rightarrow P(\mathscr{L})$ is a total function which given D and a term $t \in T = TE(D)$ returns a set of leaf vertices L of D, where L contains each leaf vertex reachable from all paths of D containing an edge labeled t, and $\mathscr{L}$ denotes the universal set of leaf webpages,
- BP (*back propagate*): $(\mathscr{D} \times P(\mathscr{L})) \rightarrow \mathscr{D}_\perp$ is a partial function which given D and L returns a DAG D', where D' contains only paths from the root of D to the leaves of D which classify the leaf vertices in L, and
- CE (*consolidate edges*): $(\mathscr{D} \times \mathscr{T}) \rightarrow \mathscr{D}_\perp$ is a partial function which given D and a term $t \in TE(D)$ returns D', where any edge e in D labeled with t is removed in D', the source v_s of e is replaced with its target v_t in D', and v_t becomes the new target of any edge e' with target v_s in D'.

[1] Some terms and definitions in this section have been reported by the author in (Perugini & Ramakrishnan, 2010) and appear here for purposes of clarity and comprehension.

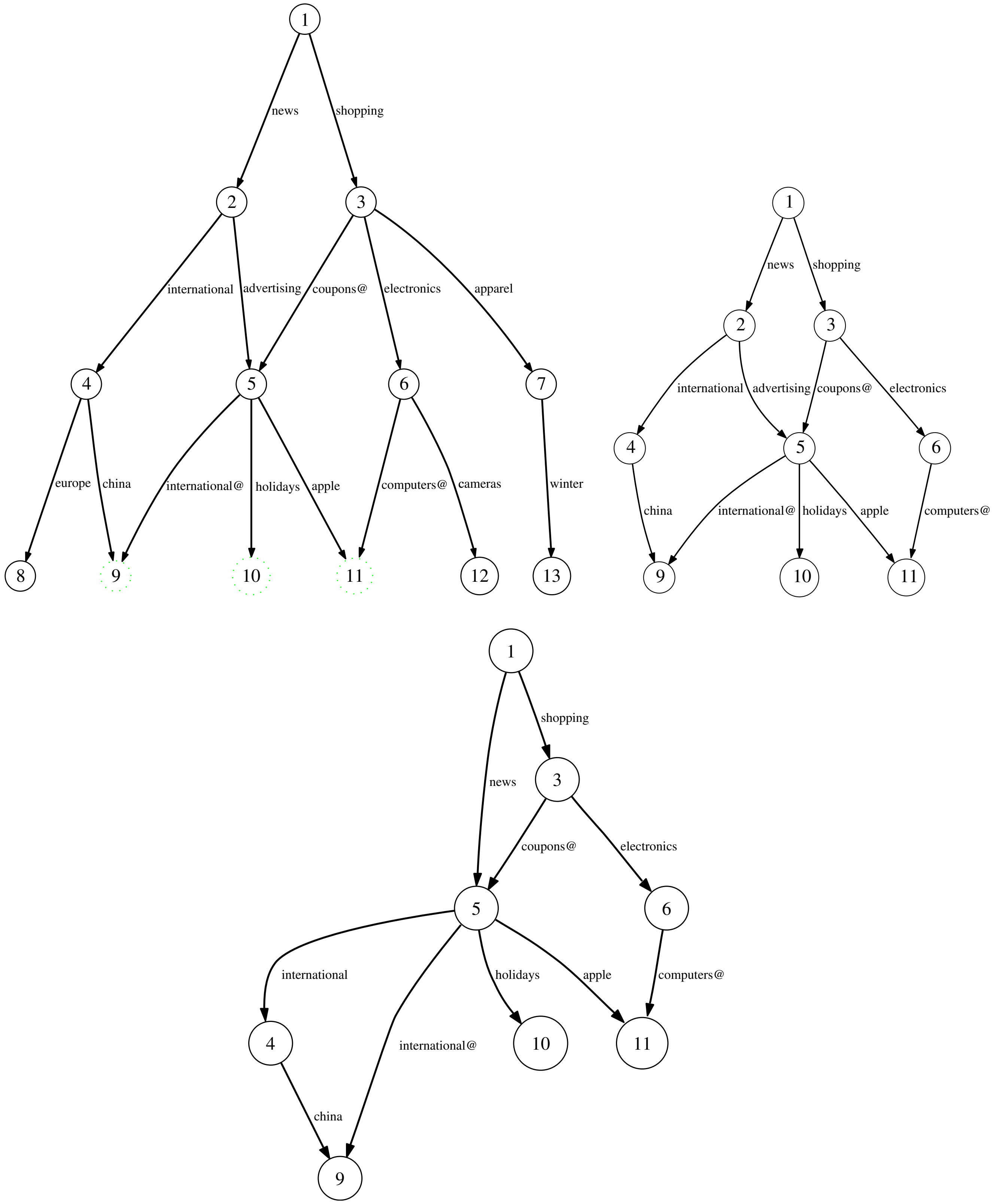

Fig. 1. Website transformations simplified for purposes of presentation: illustration of forward-propagation (FP) followed by back-propagation (BP) on the DAG on left. (left) A sample DAG model of a hierarchical website. Vertices 9, 10, and 11 (i.e., those dotted) represent the result of forward-propagation wrt the term 'advertising': $FP(D, \text{advertising})$. (center) Result of back-propagation wrt leaf vertices 9, 10, and 11 on left: $BP(D, FP(D, \text{advertising}))$. (right) Result of out-of-turn interaction with the DAG D shown on left wrt the term 'advertising': $OOT_1(D, \text{advertising})$. Alternatively, we can think of this DAG as the result of consolidating edges with the DAG D' in center (i.e., $CE(D', \text{advertising})$).

Fig. 1 illustrates the out-of-turn transformation (i.e., forward-propagation (left) followed by back-propagation (center) followed by consolidation (right)).

Intuitively, this transformation retains all sequences of D which contain the out-of-turn input (*FP* followed by *BP*), and then removes the out-of-turn input from those remaining sequences (CE). The result of *FP* is the set of all leaf vertices classified by the out-of-turn input. We back-propagate from this set of leaves up to the root of the DAG with *BP*. Note that when no term in the DAG represented by the first argument to OOT_1 resides at more than one level, and the second argument to OOT_1 is in-turn information, the transformation is functionally equivalent to B. Thus, OOT_1 subsumes B.

To marry the out-of-turn transformation with standard techniques from information retrieval we can replace *FP* with any total function *SL* (*select leaves*): $(\mathscr{D} \times \mathscr{T}) \rightarrow \mathscr{L}$ which given D and a term $t \in TE(D)$ returns a set of leaf vertices of D (*FP* is an instance of *SL*). This generalization leads to the possibility of bringing units of terminal information (i.e., terms modeled in the leaf pages and not explicitly used in the classification), in replacement of or in addition to structural information, to bear upon the transformation and resulting interaction. For instance, we might perform a query (e.g., 'laptop') in a vector-space model over the set of leaf webpages (i.e., documents) using cosine similarity to arrive at a target set of leaves from which to back-propagate. Notice that D also can be represented as a $|TE(D)| \times |CR(D)|$ term-document matrix, where rows correspond to terms (i.e., structural information, or edge-labels) and the columns correspond to webpages (i.e., terminal information, or leaf vertices). *Collect results* is a total function $CR : \mathscr{D} \rightarrow P(\mathscr{L})$ which given D returns a set of all the leaf vertices in D. For instance, $CR(D)$ returns the $\{9, 10, 11\}$ set of vertices, where D is the DAG in Fig. 1 (center).

3.3. Commutativity

We now examine the commutativity of the out-of-turn transformation.

Lemma. *The out-of-turn transformation is commutative, assuming both sides are defined:*

$$OOT_1(OOT_1(D, x), y) = OOT_1(OOT_1(D, y), x),$$

where x and y represent terms. A sketch of the proof of this lemma is given in (Perugini, 2004, Chap. 4)

Armed with this lemma, we can consider the possibility of communicating multiple terms per utterance, where an *utterance* is a set of terms with the same *arrival time* — the time at which the user communicates a term or terms to the system. To accommodate multiple terms per utterance, we re-define the out-of-turn transformation:

$$OOT(D, u) = OOT_1(\cdots OOT_1(OOT_1(D, t_1), t_2) \cdots, t_n),$$

where u denotes an utterance consisting of only the $\{t_1, t_2, \ldots, t_n\}$ set of terms and each OOT_1 on the rhs refers to (1). If $OOT(D, u)$ returns a DAG containing only one vertex v (and, therefore, no edges), then the utterance u is complete information (and v is terminal information). Otherwise, u is partial information.

3.4. Web interaction

We now present concepts which relate to a user's interaction with a website to help describe the cumulative effect of the out-of-turn transformation on a site. Several partial orders can be defined over an interaction set wrt arrival time. When a user clicks on a hyperlink, she implicitly communicates the hyperlink's label to the underlying system. For instance, when a user clicks on a hyperlink labeled 'news' followed by that labeled 'international', she communicates the $\prec$news, international$\succ$ terms to the system, in that order. Similarly, when the user supplies out-of-turn input, he is communicating terms to the system. These partial orders can be summarized as partially ordered sets or posets. Each linear extension of such a poset is a total order called an *interaction episode*. A *browsing interaction episode* of D is a total order on any interaction set of D wrt the parenthood relation of D. Notice that a browsing episode is the same as a sequence as defined above. An *out-of-turn interaction episode* is a total order over the set of all set partitions of an interaction set wrt the arrival time relation implied by out-of-turn interaction. The arrival time relation implied by out-of-turn interaction is a partial order containing only the reflexive tuples of all set partitions from any interaction set. In other words, out-of-turn interaction requires none of the term set partitions from each interaction set are required to be ordered. The linear extensions of the posets associated with these partial orders are out-of-turn interaction episodes.

An *interaction paradigm* $\mathscr{P}$ for D is the union of all linear extensions of posets defined over all interaction sets of D. In other words, an interaction paradigm is a complete set of realizable interaction episodes from D wrt a transformation (e.g., *Browse* or *OOT*). The browsing paradigm $\mathscr{P}_B$ of D in Fig. 1 (left) is:

{$\prec$news, international, europe$\succ$, $\prec$news, international, china$\succ$, $\prec$news, advertising, international$\succ$,
$\prec$news, advertising, holiday$\succ$, $\prec$news, advertising, apple$\succ$, $\prec$shopping, coupons, international$\succ$,
$\prec$shopping, coupons, holiday$\succ$, $\prec$shopping, coupons, apple$\succ$, $\prec$shopping, electronics, computers$\succ$,
$\prec$shopping, electronics, cameras$\succ$, $\prec$shopping, apparel, winter$\succ$}.

The out-of-turn paradigm $\mathscr{P}_O$ of D is:

{$\prec$(europe international news)$\succ$, $\prec$(international news), europe$\succ$, $\prec$(europe news), international$\succ$,
$\prec$(europe international), news$\succ$, $\prec$news, (europe international)$\succ$, $\prec$international, (europe news)$\succ$,
$\prec$europe, (international news)$\succ$, $\prec$news, international, europe$\succ$, $\prec$news, europe, international$\succ$,
$\prec$international, news, europe$\succ$, $\prec$europe, news, international$\succ$, $\prec$international, europe, news$\succ$,
$\prec$europe, international, news$\succ$, all permutations of all set partitions of {news, international, china},
..., all permutations of all set partitions of {shopping, electronics, cameras},
$\prec$(apparel shopping winter)$\succ$, $\prec$(apparel shopping), winter$\succ$, $\prec$(shopping winter), apparel$\succ$,
$\prec$(apparel winter), shopping$\succ$, $\prec$shopping, (apparel winter)$\succ$, $\prec$apparel, (shopping winter)$\succ$,
$\prec$winter, (apparel shopping)$\succ$, $\prec$shopping, apparel, winter$\succ$, $\prec$shopping, winter, apparel$\succ$,
$\prec$apparel, shopping, winter$\succ$, $\prec$winter, shopping, apparel$\succ$, $\prec$apparel, winter, shopping$\succ$,
$\prec$winter, apparel, shopping$\succ$},

where terms in parentheses (e.g., '(europe news)') represent a single utterance with multiple terms (i.e., more than one term with the same arrival time). Since *OOT* subsumes *Browse*, the browsing paradigm of a site D is always a proper subset of the site's out-of-turn interaction paradigm (Perugini, 2004, Chap. 4), There are 143 (= $|\mathscr{P}_O|$) interaction episodes in the out-of-turn interaction paradigm of D in Fig. 1 (left).

To capture the number of episodes in an out-of-turn paradigm we use notation from discrete mathematics (Kreher & Stinson, 1999, §3.2: Set partitions, Bell & Stirling numbers), where $s(m)$ is the set of all partitions of a set of size m into non-empty subsets (where m is a positive integer), and $s(m,n)$ is the set of all partitions of a set of size m into exactly n non-empty subsets (where n is a positive integer and $n \leqslant m$). The *Bell number* of a set of size m is $B(m) = |s(m)|$. The *Stirling number* of a set of size m is $S(m,n) = |s(m,n)|$. It follows that $B(m) = \sum_{n=1}^{m} S(m,n)$.

Intuitively, support for the out-of-turn transformation in a website enriches user interaction with that site so that users can experience traversals through the site which represent permutations of all set partitions of the interaction set from each browsing episode in the site. Therefore, we use this notation to count permutations of set partitions. Specifically, we define *size of out-of-turn paradigm* as a total function $SP_O : \mathscr{D} \rightarrow \mathbb{N}$ which given D returns the size of its out-of-turn interaction paradigm (i.e., the total number of interaction episodes in the paradigm):

$$|\mathscr{P}_O| = SP_O(D) = \sum_{E \in SQ(D)} \sum_{n=1}^{|GIS(E)|} n! \times S(|GIS(E)|, n),$$

where

- SQ (*sequencize*): $\mathscr{D} \rightarrow P(\mathscr{E})$ is a total function which given D returns the browsing paradigm $\mathscr{P}_B$ of D, where $\mathscr{E}$ represents the universal set of interaction episodes, and
- GIS (*get interaction set*) $\mathscr{E} \rightarrow \mathscr{S}$ is a total function which given an interaction episode E returns the interaction set over which it is defined, where $\mathscr{S}$ denotes the universal set of sets.

This formula considers valid utterances containing one or more terms and makes no assumption on the consistency of the length (i.e., number of terms) across all browsing episodes.

The columns of Table 1 labeled $n = 1, \ldots, 10$ contain the number of episodes (permutations) of n partitions of an interaction set of size m. The column labeled $|\mathscr{P}_O|$ gives the sum of the columns labeled $n = 1, \ldots, 10$ for a particular row or, in

Table 1
The number and type (i.e., partitioned into $n = 1, \ldots, 10$ utterances) of interaction episodes enabled by the use of the out-of-turn transformation in a site with one sequence of length m. $|\mathscr{P}_B| = 1$ in all rows.

$n! \times S(m,n)$										m	$\lvert\mathscr{P}_O\rvert$	%Δ
n = 1	2	3	4	5	6	7	8	9	10			
1										1	1	0
1	2									2	3	200
1	6	6								3	13	1200
1	14	36	24							4	75	7400
1	30	150	240	120						5	541	54 000
1	62	540	1560	1800	720					6	4683	468 200
1	126	1806	8400	16 800	15 120	5040				7	47 293	4 729 200
1	254	5796	40 824	126 000	191 520	141 120	40 320			8	545 835	54 583 400
1	510	18 150	186 480	834 120	1 905 120	2 328 480	1 451 520	362 880		9	7 087 261	708 726 000
1	1022	55 980	818 520	5 103 000	16 435 440	29 635 200	30 240 000	16 329 600	3 628 800	10	102 247 563	10 224 756 200

other words, the size of the out-of-turn paradigm corresponding to a browsing paradigm consisting of only one episode. The ratio of the number of episodes in a DAG's out-of-turn paradigm to those in its browsing paradigm is shown in the column labeled %Δ in Table 1 and defined by the following expression:

$$\Delta(D) = \frac{|\mathscr{P}_O| - |\mathscr{P}_B|}{|\mathscr{P}_B|} = \frac{SP_O(D) - |SQ(D)|}{|SQ(D)|}.$$

4. Practice

We now study the effect that the out-of-turn transformation has on two websites and, specifically, the implications of the structure and characteristics of a site on the results of the transformation.

4.1. Analysis toolkit

Given the formalism above, users can make out-of-turn utterances at any point while interacting with a website, and in any order, and, thus, there is a combinatorial explosion in the number of possible interaction episodes this single transformation supports (cf. the column labeled $|\mathscr{P}_O|$ of Table 1). To help site designers explore the interaction episodes the out-of-turn transformation enables, we built a toolkit, available from http://oot.cps.udayton.edu/oot-toolkit.tgz, which consists of Perl scripts to: (i) compute the size of an out-of-turn paradigm of a site given the number and length of each sequence therein, (ii) generate the out-of-turn interaction paradigm for a site (i.e., given the browsing paradigm, enumerate all possible interaction episodes realizable through the out-of-turn transformation), (iii) simulate interaction episodes in batch while collecting a variety of transformation statistics, and (iv) compute the number of sequences in which each term in a site is contained. The scripts to both compute the size of an out-of-turn paradigm and generate all interaction episodes in an out-of-turn paradigm make use of modules which are optimized to use dynamic programming strategies and, as a result, the toolkit produces results on large sites fast.

The episode simulator produces a complete summary of what is capable with support to interact out-of-turn with a website. For instance, on the sample DAG in Fig. 1 (left), it produces one line per interaction episode followed by a colon and the number of sequences through the site remaining after each utterance in that episode:

```
(news) (international) (europe): 11 5 3 1
(news) (europe) (international): 11 5 1 1
(international) (news) (europe): 11 4 3 1
(europe) (news) (international): 11 1 1 1
(international) (europe) (news): 11 4 1 1
(europe) (international) (news): 11 1 1 1
(news international) (europe): 11 3 1
...
(winter) (apparel) (shopping): 11 1 1 1
(shopping apparel) (winter): 11 1 1
(shopping winter) (apparel): 11 1 1
(apparel winter) (shopping): 11 1 1
(shopping) (apparel winter): 11 6 1
(apparel) (shopping winter): 11 1 1
(winter) (shopping apparel): 11 1 1
(shopping apparel winter): 11 1.
```

Notice that the numbers trailing some of the above episodes have repeating ones (1's). Once an utterance initiates a transformation which renders a site with only one remaining sequence, the result is effectively fixed. It is up to the designer to force the user to click through a series of links leading to the only terminal page remaining or to consolidate that series.

4.2. Case studies

The out-of-turn transformation is a pruning operator and is appropriate on the web when several term associations underlie the hierarchical model of the site on which it is applied. The sequences pruned from a site are those which do not contain the term supplied out-of-turn. Therefore, invoking the out-of-turn transformation with a term which appears in a frequent number of sequences results in the retention of more sequences (and, thus, terminal pages) and removal of less. On the other hand, suppling a term out-of-turn which appears only in a few sequences selects less sequences and prunes more. For instance, the term 'advertising' classifies six of the 11 sequences through the DAG D in Fig. 1 (left) and, therefore, supplying 'advertising' out-of-turn causes the site to be thinned while retaining the semblance of a hierarchy

(Sacco, 2000). In contrast, the term ‘apparel’ is occurs in only one sequence in *D* and, thus, saying it out-of-turn results in a shortcut directly to terminal page 13 (Gerstel et al., 2007). Therefore, while the definition of the transformation is fixed as shown in Eq. (1) (i.e., it is applied consistently across different sites), its results as well as the interaction afforded to the user depend on how the terms labeling hyperlinks are distributed throughout the site’s sequences.

For insight into the results of the out-of-turn transformation in practice, we conducted a variety of experiments in two websites: Project Vote Smart (PVS) and the News category of the Open Directory Project (hereafter referred to as News). Terms are distributed differently throughout the sequences of these sites. Specifically, unlike the sample site given in Fig. 1 (left), each level of PVS corresponds to a facet of information-seeking. At the first level, users are asked to make a *state* selection, followed by *branch of Congress* (House or Senate), then a choice for political *party* (Democrat, Republican, or Independent), and, finally, a choice for *district/seat*. We say such sites are *faceted* (Perugini, 2009) because each level of the site corresponds to a facet of information seeking. We call faceted sites with a consistent depth (i.e., sequence length across all sequences or number of facets across all sequences) *structured* (Perugini, 2009). On the other hand, the sample site given in Fig. 1 (left) and ODP are without a facet classifying the terms at each level. We call such sites *semistructured* because the data they present is schemaless and self-describing and, thus, often called *semistructured data* (Abiteboul, Buneman, & Suciu, 2000). Furthermore, News, and ODP in general, unlike the sample site in Fig. 1 (left), does not have a consistent depth across all sequences.

Project Vote Smart is a comprehensive and authoritative website for political officials at all levels of government. PVS has a webpage for each state and federal official containing biographical information as well as information about the official’s party affiliation, committees, and voting record. The Open Directory Project (ODP) is the largest, most comprehensive, and most widely distributed human-compiled taxonomy of links to websites (Perugini, 2008; The Open Directory Project, 2002). We analyzed the News topic of ODP available from http://rdf.dmoz.org in RDF format.

Table 2 captures values for a variety of structural characteristics of PVS and News. Results in Tables 2 and 3 reflect the US congressional landscape on December 4, 2007 and the News category of ODP data based on the `structure.rdf.u8.gz` RDF dump file, downloaded on January 18, 2008, which contains the category hierarchy information. We also include the values of these characteristics for the sample site in Fig. 1 (left) for purposes of comparison. While the site in Fig. 1 (left) is a DAG owing to the presence of symbolic links, News is not a DAG due to the presence of symbolic links which induce cycles. While so-called *hard links* create the natural parent-child relationships in a tree, the source vertex of a symbolic link is actually not the parent of its target vertex though it appears to be. Symbolic links create multiclassification in directories (Perugini, 2008) and are suffixed with @ in the ODP and *Yahoo!* directories (Perugini, 2008). The edge labeled ‘coupons@’ from vertex 3 to 5 in Fig. 1 (left) is a symbolic link. PVS is a tree since it does not use symbolic links. Since the targets of some of the symbolic links in News reside outside of the News section of ODP, we purged the symbolic links from News and analyzed a tree model of it.

Values in the column labeled **Depth** indicate the minimum and maximum length of any sequence. In Fig. 1 (left) and PVS, the minimum and maximum depth equal each other. The column labeled **#Tv** provides the sum of the values from the corresponding entries of the columns labeled **#Nlv** and **#Lv**. In the absence of symbolic links, the number of hyperlinks in a site (given in the column labeled **#Lk**) is equal to the number of child vertices in the site (or one minus the number of total vertices since the root is a child of no vertex). We define *term* as a string labeling a hyperlink (i.e., the complete text between the `<a href="">` and `</a>` HTML tags). In Fig. 1 (left), ‘news’ is a term. Notice that while each term in Fig. 1 (left) contains only one word (i.e., any string of characters except space), this definition permits a term to consist of more than one word (e.g., ‘Business and Economy’ is one term in News) and this viewpoint is reflected in the terms counts in Table 2, which omit duplicates. The number of duplicate terms is given in the column labeled **#Lk** since the total number of terms equals the total number of hyperlinks based on the definition of term above. We compute the average (μ) number of children per vertex as the total number of children (i.e., the total number of vertices minus one, divided by the total number of parents or non-leaf vertices, in a site). Table 2 reveals that there are more leaves than non-leaves in the two sites studied (1.69 times more in PVS and six times more in News). The terms in this paragraph are also defined in (Perugini, 2008).

4.3. Results

Table 3 reveals that even though News has 100 less sequences than PVS, the out-of-turn transformation enables more interaction episodes in it. The minor increase in depth in News (none of its sequences extend more than two levels deeper than any in PVS) versus PVS translate to more than four times the number of episodes. However, in both sites, support for out-of-turn interaction drastically increases the scope of ways to interact with the site (cf. column labeled %Δ in Table 3).

Table 2
Structural characteristics of the sites we studied.

Site	URL	Type	S?	Depth	#Tv	#Nlv	#Lv	#Sq	#Lk	#UT	μ C/Nlv
Fig. 1 (left)	–	DAG	×	3	13	7	6	11	15	14	1.71
PVS	http://vote-smart.org	tree	√	4	857	319	538	538	856	116	2.68
News (ODP)	http://dmoz.org/news	graph	×	[2–6]	511	73	438	438	510	292	6.99

S = structured, Tv = total vertices, Nlv = non-leaf vertices, Lv = leaf vertices, Sq = sequences, Lk = links, UT = unique terms, C/Nlv = children per non-leaf vertex.

Table 3
Statistics on the sizes of the browsing and out-of-turn paradigms in the sites we analyzed.

Site	$\|\mathscr{P}_B\|$	$\|\mathscr{P}_O\|$	%Δ	#UET
Fig. 1 (left)	11	143	1200	19
Project Vote Smart	538	40350	7400	1022
News (ODP)	438	177594	40447	1142

UTE = unique episode types.

While there is a stark difference in the number of episodes supported by the out-of-turn transformation in each site, the number of distinct episode types is nearly identical. Specifically, the 40,350 total episodes in PVS fall into 1,022 unique episode types and the 177,594 total episodes in News fall into 1,142 unique types. An episode type indicates how the episode transforms the site's structure irrespective of the content of the utterances in the episode. For instance, in News, the episodes $\prec$(Analysis and Opinion Columnists), (Directories)$\succ$ and $\prec$(Analysis and Opinion Columnists), (By Publication)$\succ$ both leave the site with 14 sequences after the first utterance and only one after the second utterance and, therefore, both episodes have the same type. Fig. 2 (top left) examines the number of sequences remaining after each utterance across all episodes in News. Each line from the maximum y value (i.e., the starting number of sequences in a site) to the minimum y value (which is always one) represents an episode. The dense areas of this graph depict the dominate episodes types.

Note that in cases such as Fig. 1 (left, $m = 3$) and PVS ($m = 4$), where every leaf in the DAG model of the site resides at the same depth, the values in the column labeled %Δ in Table 3 are the same as those in the column labeled the same in Table 1. Therefore, while the values for $|\mathscr{P}_O|$ and %Δ in Table 1 are wrt one browsing episode, they are relevant to the computation of the number of episodes in sites with a consistent length across all sequences.

To reiterate, the result of the out-of-turn transformation depend on the frequency of sequences in which the term supplied out-of-turn is contained. For instance, the term 'democrat' occurs in 286 of the 538 sequences through PVS and, therefore, supplying it out-of-turn causes the hierarchy to be thinned. In contrast, the term 'Washington, DC' occurs in only one sequence in PVS and, thus, saying it out-of-turn results in a shortcut directly to the terminal webpage of the democrat

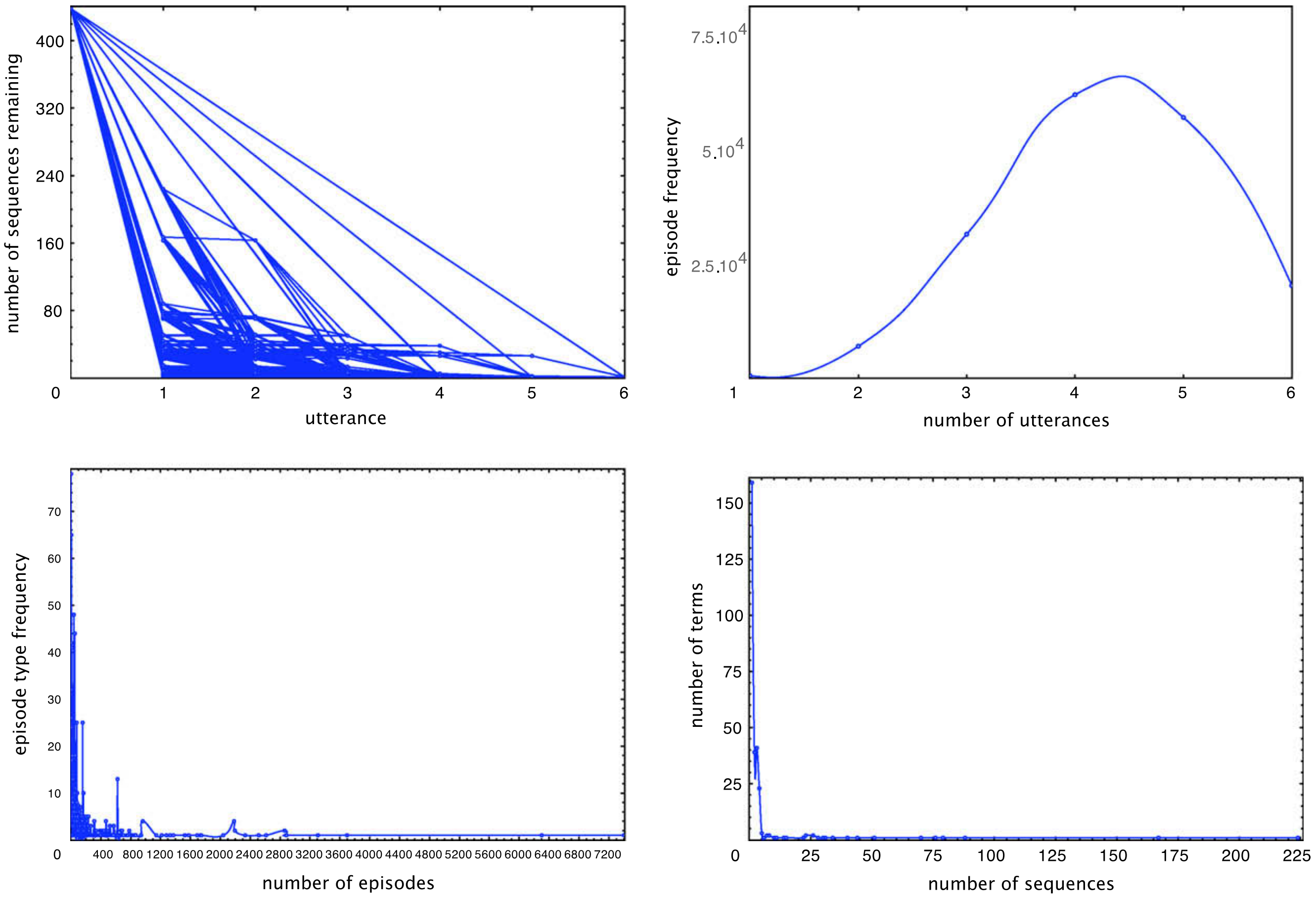

Fig. 2. Graphs to help explore the cumulative effect of the out-of-turn transformation on News: (top left) episodes, (top right) distribution of episodes across utterances, (bottom left) distribution of episode types across episodes, (bottom right) distribution of terms across sequences.

congressperson representing DC in the House. Similarly, in News, the term 'United States' appears in a high percentage of all sequences and, therefore, saying it out-of-turn results in a reduced hierarchy, while the term 'agricultural' is only resides in one sequence and, thus, saying it out-of-turn results in a shortcut. Therefore, out-of-turn interaction supports episodes requiring interaction ranging from that required by a shortcut (i.e., one utterance) to that required by browsing episodes (recall, any out-of-turn paradigm subsumes the corresponding browsing paradigm). In other words, the technique may require up to as many utterances (or steps) as the depth of each sequence to find the desired information. Therefore, to explore this continuum we classify the episodes based on how many utterances each requires of the user to arrive at the desired information. Fig. 2 (top right) provides a graph of this distribution in News and illustrates that the episodes in News appear to be somewhat normally distributed throughout this continuum.

We also plot the distribution of the episode types over the number of episodes per type in Fig. 2 (bottom left) for News. This graph illustrates that many types have a small number of episodes and only a few types have a large number of episodes. Lastly, in Fig. 2 (bottom right) we plot the distribution of terms across sequences for News. This graph reveals that only a small number of terms are contained in a high percentage of all sequences and most terms are contained in a small number of sequences (less than 10). Intuitively this means that only a few utterances are necessary to reduce the site to a small number of sequences. In other words, the transformation helps users arrive at the desired information after only a few interactions with the site. Moreover, this result relative to the number of unique terms in each site indicates that many terms do not exist in many sequences. In summary, the frequency of sequences containing each term in a website indicates how many sequences the out-of-turn transformation prunes. The purpose of the analysis toolkit is to help designers explore the effect of the transformation in a site.

5. Discussion

Interacting out-of-turn is a helpful when a user is not sure which terminal information they desire. The technique can be used for focused or exploratory tasks. A search for 'democratic senators' in PVS is focused while navigating the congresspeople through a personalized motif (by party, branch, district/seat, state vs. by state, branch, party, district/seat) is exploratory. In either case, the transformation supporting the technique vastly increases the scope of ways to interact with a hierarchical website and, thereby, accommodates multiple models of information-seeking. Overall, the cumulative effect of a series of out-of-turn transformations on a site converts the site's one-size-fits-all browsing paradigm (e.g., by-genre, by-author), insufficient to satisfy all users, into one from which all users are accommodated (e.g., by-genre, by-author *and* by-author, by-genre), albeit without explicitly enumerating all possible traversals in the hardwired hyperlink structure or building a user model for each individual user. The transformation is appropriate in:

- large, semistructured web directories, such as ODP, where it can both thin the directory when used with a term appearing in a high percentage of sequences (and preserve hierarchical structure and, thus, context) or help users identify a particular sequence in a directory given a term which appears in only a few sequences or only one sequence resulting in a shortcut. In both cases, it helps the user identify where the desired information is situated within a voluminous directory,
- faceted sites where there are too many facets of information seeking to render using a relational table interface, such as Apple's *iTunes* or the *Relational Browser++* (Zhang & Marchionini, 2004), for purposes of limited screen real estate, especially in systems designed for mobile devices (e.g., PDAs, smart phones), and
- automated, interactive voice-response systems such as those used by banks and airlines to reduce agent costs in call routing (Perugini et al., 2007).

We contribute an alternate perspective on website transformation and its cumulative effect on personalized user interaction, a theoretical model for site transformation, an analysis toolkit to investigate the effects of the out-of-turn transformation on site structure, and two case studies, which involve experiments that demonstrate the capabilities of the transformation and toolkit. Our approach lies between template- and AI-based (i.e., mining user logs to build user models of preference) approaches to web personalization.

Acknowledgments

My view of personalization and website transformation has benefited from interactions with Naren Ramakrishnan at Virginia Tech. I also thank Ed Fox at Virginia Tech for helpful discussions and insights. This work was supported in part by seed grants from the University of Dayton Research Council, the Ohio Board of Regents, and the University of Dayton College of Arts and Sciences.

References

Abiteboul, S., Buneman, P., & Suciu, D. (2000). *Data on the web: From relations to semistructured data and XML*. San Francisco, CA: Morgan Kaufmann.
Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-adapted Interaction, 11*(1–2), 87–110.
Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology, 3*(1), 1–27.
Gerstel, O., Kutten, S., Laber, E., Matichin, R., Peleg, D., Pessoa, A., et al (2007). Reducing human interactions in web directory searches. *ACM Transactions on Information Systems, 25*(4), 1–27. article 20.

Hearst, M. (2000). Next generation web search: Setting our sites. *IEEE Data Engineering Bulletin, 23*(3), 38–48.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Finding the flow in web site search. *Communications of the ACM, 45*(9), 42–49.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to usenet news. *Communications of the ACM, 40*(3), 77–87.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM SIGKDD Explorations, 2*(1), 1–15.

Kreher, D., & Stinson, D. (1999). *Combinatorial algorithms: Generation, enumeration, and search*. Boca Raton, FL: CRC Press.

Manber, U., Patel, A., & Robinson, J. (2000). Experience with personalization on Yahoo! *Communications of the ACM, 43*(8), 35–39.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM, 43*(8), 142–151.

Narayan, M., Williams, C., Perugini, S., & Ramakrishnan, N. (2004). Staging transformations for multimodal web interaction management. In M. Najork & C. Wills (Eds.), *Proceedings of the thirteenth international ACM world wide web conference (WWW)* (pp. 212–223). New York, NY: ACM Press.

Perkowitz, M., & Etzioni, O. (2000). Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence, 118*(1–2), 245–275.

Perugini, S., 2004. Program Transformations for Information Personalization. Ph.D. dissertation, Department of Computer Science, Virginia Tech, available in the Virginia Tech ETD collection at <http://scholar.lib.vt.edu/theses/available/etd-06252004-162449/>.

Perugini, S. (2008). Symbolic links in the open directory project. *Information Processing and Management, 44*(2), 910–930.

Perugini, S. (2010). Supporting multiple paths to objects in information hierarchies: Faceted classification, faceted search, and symbolic links. *Information Processing and Management, 46*(1), 22–43.

Perugini, S., Anderson, T., & Moroney, W. (2007). A study of out-of-turn interaction in menu-based, IVR, voicemail systems. In M. Rosson & D. Gilmore (Eds.), *Proceedings of the twenty-fifth international ACM conference on human factors in computing systems (CHI)* (pp. 961–970). New York, NY: ACM Press.

Perugini, S., & Ramakrishnan, N. (2003). Personalizing interactions with information systems. In M. Zelkowitz (Ed.). *Advances in computers* (Vol. 57, pp. 323–382). Amsterdam: Academic Press.

Perugini, S., Ramakrishnan, N., 2010. Program transformations for information personalization. Computer Languages, Systems and Structures, doi:10.1016/j.cl.2009.09.002.

Riedl, J. (2001). Personalization and privacy. *IEEE Internet Computing, 5*(6), 29–31.

Sacco, G. (2000). Dynamic taxonomies: A model for large information bases. *IEEE Transactions on Knowledge and Data Engineering, 12*(3), 468–479.

Sacco, G., & Tzitzkas, Y. (2009). *Dynamic taxonomies and faceted search: Theory practice and experience* (Vol. 25). Berlin: Springer.

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

The Open Directory Project. (2002). About the open directory project. <http://dmoz.org/about.html> Retrieved 11.12.09.

White, R., Jose, J., & Ruthven, I. (2006). An implicit feedback approach for interactive information retrieval. *Information Processing and Management, 42*(1), 166–190.

Zhang, J., & Marchionini, G. (2004). Coupling browse and search in highly interactive user interfaces: A study of the relation Browser++. In H. Chen, H. Wactlar, C. Chen, E.-P. Lim, & M. Christel (Eds.), *Proceedings of the fourth international ACM/IEEE-CS joint conference on digital libraries (JCDL)* (pp. 384). New York, NY: ACM Press.

Saverio Perugini is an Assistant Professor in the Department of Computer Science at the University of Dayton. His research interests include information personalization, web mining, and functional programming. He has a Ph.D. in Computer Science from Virginia Tech.