Genre analysis of bookmarked webpages

Michela Montesi
Faculty of Information Science, Complutense University of Madrid, Madrid, Spain

Abstract
Purpose – A total of 17 user-compiled collections of webpages, comprising 833 bookmarked links in terms of genre, are studied. The purpose of this paper is to find out whether users tend to bookmark certain web genres more than others. Genre theory helps to make sense of the different pages included in these collections, and to classify them, according to their communicative purpose and salient non-topical features, into blogs, search interfaces, articles, tutorials.

Design/methodology/approach – A total of 17 participants took part in the research by providing their collections of bookmark links. They were also interviewed about the reasons for bookmarking and to comment on their collections. Relying on the interview results and on the previous literature, the bookmarks were classified into four super-genres: main or access pages, transactional pages, navigational pages, and content pages.

Findings – The results of the classification into web genres revealed a clear tendency to bookmark main pages, such as homepages, which accounted for 42 per cent of all bookmarked web links. Moreover, some aspects of relevance were highlighted such as the connections to use, time, and context, as well as to the main web activity (browsing or searching).

Originality/value – Previously, bookmarks have mostly been studied as tools for information reuse, but very rarely as sources of implicit relevance feedback. In addition, from the point of view of genre theory, this research shows the importance of relating web genres to users’ intentions behind queries.

Keywords Worldwide web, Information retrieval

Paper type Research paper

Introduction
Relevance is a fundamental notion in Information Science. Whereas cataloguing and classification are concerned with aboutness, relevance is a central principle of searching and thus of information retrieval (IR) (Saracevic, 2007b). Though the concept is intuitive, its implementation in working IR systems is a challenging research topic for information science, especially when it comes to searching the web. As relevance is a contextual phenomenon which is meaningful for a certain searcher, information need and task, a considerable amount of research into relevance deals with searchers or information system users, and with how their feedback might improve IR systems, either explicitly or implicitly. Kelly and Teevan (2003) review research into what is called “implicit relevance feedback techniques” for expansion of web queries, i.e. techniques that “unobtrusively obtain information about users by watching their natural interactions with the system” (Kelly and Teevan, 2003, p. 18). The time spent reading a webpage, or actions such as clicking, printing, saving, or selecting, all convey implicit feedback about user relevance judgements. Implicit feedback data cannot be considered as reliable as data collected with explicit feedback techniques, when the user specifically indicates whether an item is relevant or not, and to what extent. However implicit data can be gathered easily, in large amounts, and with no hassle for
the user (Jung et al., 2006). Research into explicit relevance feedback, conversely, pretends to define the criteria users employ to make relevance decisions, in order to come up with a set of more or less generalisable criteria useful for improving IR systems. Major results show that a wide range of criteria are used to select relevant documents on the web, though topicality seems to be the most common. A certain degree of agreement exists between different studies on the relevance criteria most often relied on, but it has been proven that their relative importance changes according to the circumstances. Kim and Oh (2009) for example found that the social dimension of a Q&A site has an impact on the relevance criteria adopted in this environment. Crystal and Greenberg (2006), studying the behaviour of health information seekers, conclude that relevance judgements are complex and multifaceted and, among other things, change depending on whether the user looks at the full text document or at a surrogate. More importantly they state that relevance criteria, being closely tied to users’ situations and information needs, are potentially numerous.

Bookmarking is included among the various manifestations of implicit relevance feedback: by bookmarking a web document, users are assumed to indicate that it responds, in total or in part, to a specific contextual information need. Nonetheless bookmarking has been studied very little in comparison with other sources of implicit relevance feedback, and more often as a strategy for information reuse than as an implicit recognition of relevance. From the perspective of information reuse, bookmark files are defined as “personal web informationspaces” (Abrams et al., 1998), that, among the various methods for re-use (such as the browser history or sending emails to oneself), seem to score poorly in portability, number of access points, and ability to work as reminders (Jones et al., 2001). From the point of view of relevance, bookmarking is an intentional action bound to a specific search and context, and thus potentially revealing of the reasons that lead to establishing relevance. In this research paper, bookmarks are studied as sources of implicit relevance feedback. The way we look at them is different from the type of analysis normally carried out when studying implicit relevance feedback, as in this case we are comparing saved/bookmarked items with other items within a collection of relevant web pages. In order to do so, we analyse bookmarks from the perspective of genre theory, looking for genre regularities in the web documents bookmarked during search.

Genre theory is the background of research conducted in several different disciplines, including discourse studies, computational linguistics, organisational studies, information science, and information retrieval. Though they pursue different objectives, they all share a common theoretical background, i.e. the idea of genres as typified rhetorical actions occurring in recurrent situations (Miller, 1984). The concept was initially conceived of in literary and discourse studies, Bakhtin’s (1986) essay on speech genres probably being one of most influential formulations. More recently genre has become a widely studied subject in different branches of discourse studies, including applied linguistics (Swales, 1990) and computational linguistics (Biber, 1989). In the social sciences its popularity has risen after its application to narratives and communication patterns within organisations (Yates and Orlikowski, 1992). The potential of genre to describe and classify discourse is probably one of its most attractive features, the other being its ability to capture the contextual elements of the communicative actions it conveys. Different contexts ask for different genres, just as preparing food requires a recipe, and even when the same document type is used in different contexts, its purpose and generic nature may change. Østerlund (2006) shows how a patient’s flowsheet changes its function up to six times in the context of a hospital,
serving sometimes for checking the evolution of the patient, others for administrative and service purposes, or for comparative purposes with other patients. As for the use of genre in classification, library and information science has long relied on genre as a classifying tool for films and literary fiction. The *Moving Image Genre-Form Guide* is a system of over 150 genre terms for subject access to films and television works (Yee, 1988), whilst *Genreflecting* is a reader advisory guide to popular fictional genres which has been published periodically since 1982 (Rosenberg, 1982). More recently the ability of genre to improve retrieval from large collections of documents including the web has made it of considerable interest for IR and automatic classification. Genre theory informs one of the two approaches to automatic webpage classification, complementing classification based on content or subject (Choi and Yao, 2005). One of genres’ major potentials in webpage classification is that they are especially suitable for collections made up of different types of documents. In terms of relevance we could say that genre deals with relevance criteria other than and complementary to topical criteria.

For the last decade or so, research into automatic classification of web genres, i.e. genres available on the web, such as blogs, FAQs, or e-mails, has dealt primarily with fundamental issues such as: Can genre-based classification really improve subject-based classification? What are the most appropriate machine learning model and feature sets? (Lee and Myaeng, 2004; Lim *et al.*, 2005; Kennedy and Shepherd, 2005). If these still remain problems on the web genre community’s agenda, research has moved on and other issues have become important. A relationship has been found between genre and subject, as the web genres dealing with the same or similar subjects seem to be more easily recognised as members of a genre class by automatic classifiers (Finn and Kushmerick, 2006). Additionally it has been observed that different genres may require different feature sets, and that this should be taken into account when evaluating classifiers (Kim and Ross, 2008). In general these results show the benefit of focusing web genre research on more specific domains, in which genre feature sets are devised specifically, and the relationship between genre and subject is taken advantage of.

At present research into web genres is concerned with technical issues more than anything else, as the area is still relatively young. Among other things, a standard for building web genre corpora for testing and experimentatation is still lacking (Santini *et al.*, 2010, pp. 14-19). However as automatic recognition of web genres is supposed to support web users’ navigation, the importance of users and user communities is being more frequently recognised with a focus on more specific domains, and users may be involved in research projects as labellers or classifiers of collections of pre-selected web genres, as in studies by Kim and Ross (2007) and Santini (2008). Other researchers opt for selecting web genres from specific domains, such as the academic domain, and then having users label them (Rosso, 2008). Finally web genres have proved effective in documenting web users’ information behaviour in specific contexts (Montesi and Navarrete, 2008).

Considering the potential of genre for classification and for documenting web users’ behaviour, genre appears to be an appropriate tool for making sense of bookmarked webpages. In this paper we studied the bookmark collections of 17 web users recruited at the Faculty of Information Science of the Complutense University of Madrid (Spain) from its students, administrative and service staff, and faculty. The purpose was to find out if a relation exists between web genres and bookmarked links, and if there is a tendency to bookmark certain web genres more than others. If there is a recognisable pattern in the web genres bookmarked, this may shed some light on the relevance criteria used beyond topicality that we understand are conveyed by genre. In addition
indications can be obtained regarding the web genres that users tend to consider more often relevant, which would benefit research into web genres.

Methodology
In order to determine generic patterns in the collections, we combined a manual genre classification of bookmarked webpages with a series of interviews. At the end of the interview, participants provided a copy of their folder of bookmarked webpages. They all used a standard browser, either Internet Explorer or Mozilla Firefox. The corresponding bookmark folder, sometimes organised into several thematic folders, contained the links to those webpages that the users had saved intentionally as favourites or bookmarks at different points in their search for information. We counted and classified only these links as bookmarks, rather than other webpages such as, for example, those that the browser might record as frequently visited, or the default links to help and similar pages, that both Internet Explorer and Mozilla Firefox automatically include in these folders. All folders and the links contained in them were saved on a portable device and assigned a code for each participant. Considering the personal content of these files, participants were allowed to remove those links they did not want to disclose. As we intended this process to be as private as possible, we did not ask how many links were removed.

In total 21 participants agreed to take part in the survey, though two were excluded from the research for not having a comparable set of bookmarked webpages, and two more withdrew after the interview. Consequently this report presents results from 17 participants in total: seven administrative and service staff, five faculty, and five students. The administrative and service staff included people working for different branches of the administration, technicians, and librarians. The faculty all had teaching and research duties in different specialties of information studies. Finally the five student participants were all studying at a Master’s (three) or PhD level (two) in library and information science.

Interviews
The interviews were short private conversations with each of the participants lasting between 8 and 16 minutes. They were all recorded but one, at a participant’s request, and transcribed between 25 March and 7 July 2008. The main purpose was to provide a scenario (the participants, their activities, interests and principles) to use in the following classification of webpages. During the interviews the participants first commented about the collection in general, the frequency with which they bookmark web items, and the reasons for doing so. Then they chose and focused on some of the bookmarked webpages (or groups of webpages organised in folders) explaining the circumstances in which they made the decision to save these web links, the way they used them, and other details. The interviews took place in the researcher’s or the participant’s office. The location of the interviews affected the nature of the bookmarked webpages. Administrative and service staff were all interviewed at the workplace, whereas students brought a copy of their bookmarks file from their home PC. In this way students were more likely to supply links relating to hobbies, spare time, and in general to the personal sphere of their lives. Three faculty members were interviewed at their office and provided the links stored in their work PC, whilst two others, working at different locations, sent the bookmarks via e-mail.
Classification of bookmarks into web genres

In total the collections of bookmarked items contained 1,026 webpages, of which 833 could be coded, as reported in Table I.

Links leading to error, expired, or “not found” pages were not coded and excluded from the results. A remarkable finding shown in Table I is the great differences existing among the participants at the group level and as a whole. The faculty group presents the highest median and mean of links per participant, though also the greatest average standard deviation, indicating the low generalisability of the results obtainable from the sample. A few individuals may have so strong an influence that results reflect their behaviour more than the behaviour of the group. At the group level, students present more consistent data and a much lower deviation from the mean.

Of the 17 bookmark collections provided, nine had been organised into thematic folders and the rest had not. The participants’ behaviour regarding the organisation of the collection was uneven. They might have thematic folders and at the same time keep a few loose links that are not assigned to any of them. There is an almost obvious tendency for larger collections to be split into thematic folders, which is not always true. The tendency to create folders is higher among students (four out of five had at least some), and lower among administrative and service staff (two out of seven). Faculty also tend to organise bookmarks into folders (three out of five).

The classification of web genres was carried out by a single researcher. Basically in order to classify the webpages, we relied on a previous study and the criteria applied there for classification (Montesi and Navarrete, 2008), the information gathered in the interviews, as well as previous relevant literature. Theoretically researchers have not yet agreed on an undisputable definition of genre, and Santini et al. (2010, p. 8) question whether it is needed for empirical studies and computational applications. Consequently we decided to adopt a simple practical understanding of genre, categorising each link as one web genre. Webpages are complex types of documents that can combine different textual features and functions, and a multi-genre labelling may be more appropriate to capture this complexity (Santini, 2008). However with over 800 links to classify, we opted for single-genre labelling in order to obtain a clearer picture of the whole that might show patterns in genre distribution across the 17 collections. Ambiguous cases were solved within the context of each collection, while considering all sorts of information provided by the users. Both the interviews and the labels assigned by some users to their bookmarks provided important indications of what aspects to consider when coding the webpages, though users’ labels were more concerned with the name/subject of the resource than its genre. Consequently a few genres could not be classified and were coded as unclassifiable.

<table>
<thead>
<tr>
<th></th>
<th>Total links provided</th>
<th>Expired – not coded</th>
<th>Total links coded</th>
<th>Median</th>
<th>Mean</th>
<th>Average standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin. staff</td>
<td>291</td>
<td>57</td>
<td>22</td>
<td>41.6</td>
<td>35.67</td>
<td>22</td>
</tr>
<tr>
<td>Faculty</td>
<td>551</td>
<td>129</td>
<td>97</td>
<td>110.2</td>
<td>73.44</td>
<td>97</td>
</tr>
<tr>
<td>Students</td>
<td>184</td>
<td>7</td>
<td>45</td>
<td>36.8</td>
<td>12.24</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>1026</td>
<td>193</td>
<td>41</td>
<td>60.35</td>
<td>47.91</td>
<td>41</td>
</tr>
</tbody>
</table>

Table I.

Data about the bookmark collections and links per participant
The result was a classification into genres of the 833 bookmarks. The classification resulted from several rounds of tagging in which we tried to be specific and exclusive at the same time. We tried to be specific when dealing with web genres that somehow defined the user groups, such as “library catalogues”, or with web genres better described in the literature and more easily recognisable, such as reviews or resource pages. By “resource pages” we mean “an organised set of outbound links on a given topic” (Hendry et al., 2006, p. 806). We also applied strategies to contain the number of web genres within reasonable limits that would allow us to obtain specific enough results. Some web genres were combined under general labels, such as “info pages” or “teaching material”, as explained later. Previous classifications of web genres show that the number of genres may be as high as one genre for every six webpages, even in small collections (Rosso, 2008). Rosso (2008) for instance, obtained a list of 18 web genres from a manual classification of 102 webpages, pseudo-randomly collected from the.edu domain. This and similar experiences indicate that a too fine-grained classification of web genres might fail to significantly improve subject-based classification of webpages, simply because there are far too many different web genres, corresponding variations, and names. However in limited domains or in other environments than the web, it may be highly revealing and useful.

Applying the strategies commented on above, we still obtained over 40 different web genres, which made it difficult to get a clear indication about the generic distribution of the whole collection. In order to overcome this problem we further handled the data at a super-genre level, grouping together all web genres under four general categories. For this purpose we took into account previous work on the seeking behaviour of web users, web genres, and web usage. The classification of intentions behind queries in internet search engines was especially relevant. The notion of intention is implicit in the concept of genre as a communicative action, and intentionality (objectives, roles, expectations, and motivations) is one of the aspects making up the complex concept of relevance (Saracevic, 2007a). We can assume that certain query intentions are more likely to be answered by certain web genres than others. Broder (2002) and Rose and Levinson (2004) classify intentions behind web queries into three groups: navigational, informational, and transactional/resource, depending on whether the user’s goal is to find a known page, to learn something, or to further interact or obtain a resource. This framework was rephrased and adapted to our corpus, so that we could list all web genres under one of four general categories: main pages, navigational pages, transactional pages, and content pages. “Main pages” are defined as access pages, such as homepages. Some of them cannot be regarded as proper homepages, whatever type we might choose: corporate, personal, or organisational (Shepherd et al., 2004), though they still have the function of working as a main access point, as reflected in their URL. We assume that “main pages” are the genres most likely to answer the navigational queries described by Broder (2002), i.e. looking in a search engine for something already known. By contrast, in this study, the concept of “navigational” is understood in a different and more specific way than by Broder (2002) and Rose and Levinson (2004). Here it refers to all webpages (indexes, resource pages, and tables of contents, among others) whose main function is facilitating browsing and navigation up to a content page. Navigational pages help users find their way around in the web, which is partly what main pages do as well (Montesi and Navarrete, 2008). However navigational and main pages occupy a different position in a hyperlinked structure which may account for different functions, and besides with navigational pages the main users' activity is browsing. An additional
difference is that homepages tend to have a bit of everything (indexes, search interfaces, user log-ins, advertisements, or welcoming information). For these reasons we counted them separately. The “transactional” super-genre category includes all interfaces delivering a tailored page after the user has entered some data, and download interfaces. Rose and Levinson (2004) use the term “resource” for web queries pursuing a resource. However the term “transactional” avoids a possible confusion with “resource pages”, which we regard as a type of navigational page with “an organised set of outbound links on a given topic” as in Hendry et al. (2006, p. 806). Finally the term “informational” was changed to “content” following Cooley et al. (1999). Cooley et al. (1999) differentiate between auxiliary and content pages. Auxiliary pages facilitate browsing up to the desired information, whereas content pages actually contain such information. Grouping web genres together in this way, we obtained an impressionistic picture of genre distribution in the collection of bookmarks.

More detailed variations within the four super-genre categories are also presented. The denominations assigned to the web genres, at this level of analysis, proceed from different sources, including the interviews, the information provided by the page and its URL address, and previous research. In this sense the classification reflects the nature of the corpus under scrutiny, and can be biased by the evaluator’s cultural background and the previous research available. Some categories, such as “info pages” or “indexes” are quite general, partly due to the need to pull certain genres together, as explained earlier, but also due to being unable to specify them any further. “Info pages” was broken down into “info page data”, “info page text”, and “info page advice”. “Info page data” contains such different things as statistics, weather forecasts, or stock exchange data; “info page text” usually contains informal texts that did not show any clear generic features to the evaluator; and finally “info pages advice” reports on all sort of advice, from tourist activities to cooking tips. The term “index” – lists of inbound links within the same site leading users to specific content – includes all pages of this type that could not be classified more specifically such as “table of contents” or “product catalogue”, for example. The category “other” mostly comprises unique instances, such as a list of quotations or a URL validation interface. “Teaching material” is another general denomination. In the beginning we described all teaching material as specifically as possible, because we considered that it was highly specific for the faculty and students group: exercises, course programmes, individual lesson programmes, and bibliographies. However eventually we pulled all these genres together under “Teaching material” because their individual incidence within the genre palette was almost insignificant.

**Ethical issues**

Studying collections of bookmarked webpages, we are handling personal data, which means that ethical issues come to the fore. Before they took part in the survey, participants were given a leaflet which informed them about the research, its objectives, and the measures to be taken in order to guarantee the confidentiality of all data provided. Participation was based on informed consent, and all identities were kept anonymous. For this reason no indications are given about the specific role the participants play within the administration or in the faculty in general. The UCM Faculty of Information Science is quite small and more specific information about the participants would make it easy to guess their identity. Participants were free to withdraw from the research whenever they wanted to, and to retain all items they did not want to make public from their lists of bookmarks.
Results

Limitations of the results

Before presenting the results of the classification, some observations about its limitations and scope are needed. First the corpus itself does not allow us to draw any generalisation about the sample of participants involved, as the data studied had high average standard deviations. In addition these results may be valid only within the specific context from which they proceed and may not apply to others. The study is basically explorative and results must be taken as indications for future research and theory formulation. An additional limitation has to do with the way data were manipulated, as the classification was carried out by only one coder. Other sources of bias influencing classification are commented on in the method section, such as the literature available and the nature of the corpus studied.

A different set of limitations deals with the ability of bookmarks to reflect actual users’ behaviour. Bookmarked webpages may provide less realistic depictions of web user behaviour than other data sources such as the browser history or logs, because they record only some of the users’ intentional actions. The interviews showed that links are bookmarked for a limited range of reasons, though not only for re-use, as normally accepted in studies of bookmarking. However whatever the reason, bookmarking is a user’s conscious action, which casts light on the relation between web documents and relevance judgements. Another important consideration is that bookmarked links are heavily intertwined with the context in which they are found, saved, and used, in several ways. In fact they can change from one PC to another. Some participants stressed that the bookmarks stored in their work PC were different from those in their home PC. The results must be read taking all these limitations into account. In what follows we will take a more careful look first at the information that was obtained from the interviews, and then at the classification into web genres.

Results from the interviews

The main objective of the interviews was to create a context for the classification into web genres, and to ascertain the circumstances under which web users make relevance decisions. All participants commented that the frequency with which they bookmark a webpage varies: periods of inactivity might be followed by periods of intense search and bookmarking in which many web links are collected within a short time. The consequence is that their collections change constantly; links are added, or cancelled if they are no longer interesting. Often the bookmark collections grow as links no longer of interest are forgotten and kept.

The interview results are summarised along the following three themes: how users find the webpages they bookmark, why they save them, and how they use them.

How do users find the webpages they bookmark? Bookmarks are usually found through search engines, even when the webpages they are looking for are already known. However some links can be bookmarked when they are discovered through different channels, such as during a lecture, reading printed material such as articles, listening to the radio, on recommendation, or because of an official communication, such as those coming from the central university administration for employees. Serendipity also plays an important role in discovering interesting and relevant pages worth a bookmark. One of the participants, for instance, found out about an automatic translator when using somebody else’s computer. Once back at her PC, she added it to her bookmark collection. On other occasions while searching, users may discover
webpages that have little relevance for that search but look interesting and may be of future use. These pages are bookmarked anyway.

Why do users bookmark webpages? The main reason why users bookmark webpages is their uniqueness. The term unique covers a wide range of meanings. First it means of the desired quality. Searching through the web, the amount of information users encounter is huge and often below the expected level of quality, whatever one might understand to be quality. Depending on the circumstances and the user, it means immediacy, ease of use, clarity, functionalities, completeness of information, ability to personalise the information, etc. Second unique means rare. A participant, for instance, said that he saved a newspaper webpage at a time when there were few available. Finally the concept of uniqueness has strong emotional and ideological connotations. Users may bookmark relatives’ webpages, pages with a political orientation they feel close to, webpages that look funny or deal with one’s favourite novel, etc. The importance of the emotional factors involved in bookmarking was highlighted by another factor: some participants were unable to explain in words the reasons why they had included some pages among their favourites.

A second important reason for bookmarking is that bookmarks save time, as they allow users to immediately access the resource they need. This is especially important when the resource is used daily, even if for a short period of time (for a specific assignment, to learn to use certain software, etc.). The decision to bookmark a webpage may happen after having searched for it several times in a search engine, especially if it is placed deep down in the hierarchical structure of a website.

Bookmarks are also used as reminders. As we saw, many participants get to know certain webpages when they work or search in environments other than the web. Bookmarking in these cases can be seen as a strategy to remember something. Bookmarks also work as reminders when participants are working at their PC and are engaged in something else, such as talking or chatting: being busy, they may find it useful to bookmark webpages that, for instance, their colleagues recommend to them, so that they can explore it later more quietly. Finally a webpage URL might be difficult to remember because it is too long, or it is not clear which domain it belongs to:.com or.es, etc.

How do users use the webpages they bookmark? The intended use of a bookmarked webpage is related to the concept of genre, as different genres serve different communicative or informative purposes. What the participants said about the use they make of the bookmarks was especially valuable in the classification of the bookmarks into web genres. When participants bookmark a link, they usually intend to use it for one of the following purposes:

- to support browsing, for instance when they start exploring a website;
- to compare or have access to several pages at the same time;
- to remember good pages or the search history;
- to monitor possible changes and content updates, for example of a legal nature;
- to add content regularly, such as in forums; or
- to carry out some type of transaction.

This last use was especially stressed by some participants from the administrative and service group. They bookmark pages in order to carry out specific work tasks, such as
working with specific pages for several types of operations: purchasing (books, flights [...]), control of personnel, or searching for specific data, such as ISBN codes.

**Classification of webpages into web genres**

Figure 1 shows the super-genre distribution across the 17 collections of bookmarks. Main pages are the most common genre (42 per cent), followed by content (26 per cent).

Comparing the three user groups at a super-genre level, differences among groups are observable (Figure 2). This comparison must be taken with caution considering the differences among the three sub-corpora commented on earlier.

The four super-genre categories appear in all genre palettes, including the non-classifiable items category, though their relative importance changes for each user group. The faculty and student groups present the most similarities, with most web genres falling into the “main pages” category (44 per cent and 52 per cent respectively). The administrative and service staff palette presents a strikingly different picture, with content pages leading the classification (38 per cent), followed by main (29 per cent), transactional (16 per cent), and navigational pages (15 per cent).

The multi-functionality of homepages can explain the popularity of main pages especially in the students’ palette. Student participants commented that they can use homepages for many different purposes. For instance the faculty homepage was bookmarked to directly access the virtual learning environment, to keep an eye on announcements and news, and as an index to link to the library catalogue or other resources. This multi-functionality may reach such an extent that participants do not remember homepages as such but as other genres, such as a log-in page if the main reason for saving the link is accessing one’s personal account. As for the administrative and service staff palette, routine activities can explain the considerable presence of content pages and transactional pages, which stand out compared to the other two groups (38 per cent and 16 per cent). In the interviews participants described some webpages being bookmarked for a short time, and others as more permanent links. Indeed bookmarks often, though not always, have a limited duration which ties them to the circumstance in which they were discovered.

![Figure 1. Super-genre distribution](image)
Figure 2.
Comparison of the three genre palettes at a super-genre level
and saved. After a while they may lose relevance, though it may take some time before they are deleted, if they are deleted at all. Among the most stable pages within their collections, participants in all three groups often mentioned transactional pages. It is difficult to explain the role of navigational pages in the three palettes. They represent respectively 15 per cent, 21 per cent, and 17 per cent of the administrative and service, faculty, and student palette, but were scarcely commented on during the interviews.

The composition of the super-genre categories at a more detailed level for the three user groups as a whole is shown in Table II.

The super-genre “content” presents more internal variation than any other, followed by transactional pages. Transactional pages show more internal variation than main pages, despite the size difference between the two samples: 100 items versus 346 items.

**Implications of the results and discussion**

The results of the research have implications both for relevance and for web genre research. First we will discuss the implications of the results for relevance, then for web genres, though these are necessarily interconnected. In a recent article about the concept of relevance Hjørland (2010) contends that the traditional difference established between system-relevance and user-relevance does not really exist, as both types of relevance are subjective and humanly defined. Machines cannot have tasks and goals, which are intrinsic to relevance; humans do have tasks and goals which determine what is relevant and what is not. In a certain sense this study confirms that such a dichotomy is purely conventional, because it shows that bookmarking, considered as a source of implicit relevance feedback, is embedded in a complex intertwining of phenomena. It happens in real life and not within a system or exclusively through a user-system interaction.

Bookmarking is influenced by all sorts of relationships with the context within which users access the internet, and this context includes other people, other channels, tasks, activities, emotions, etc. This is nothing new, but explains to a certain extent the reluctance to study bookmarking to derive implicit relevance feedback that we noted in the introduction to this paper: they imply such a complex net of interactions that they could not be easily translated into a machine language.

Two other aspects of relevance in connection to bookmarking emerge from this study and they concern use and time. On the one hand from the point of view of use or utility, relevance appears to be a time-bounded concept. Some web genres are mostly intended for long-term interaction, especially transactional pages, whereas others are mostly intended for short-term consumption. On the other hand bookmarking has often been seen as a strategy for information re-use, as a highly focused purpose-oriented activity. However this study shows that sometimes webpages are bookmarked without a specific task or use in mind, but simply because they mean something to the user: they are beautiful, well designed, connected to loved ones, funny, or in tune with one’s ideals or political views. Although the role that emotions play in information searching has been well known since Kuhlthau’s work (1988), it is hardly ever highlighted in connection with relevance. Hjørland (2010) for example, speaks of socially shaped factors as a determinant of relevance, but does not allude to the role of emotions or non-rational factors. At most the literature on relevance has highlighted that users are inconsistent when they judge relevance, though it has not given possible reasons (Saracevic, 2008). By looking at relevance from the point of view of bookmarking, human connotations related to the emotional sphere of existence come more clearly to light.
Table II. Composition of the three genre palettes at a micro-level

<table>
<thead>
<tr>
<th>Genre Palette</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content pages</td>
<td>219</td>
</tr>
<tr>
<td>Transactional pages</td>
<td>100</td>
</tr>
<tr>
<td>Main pages</td>
<td>346</td>
</tr>
<tr>
<td>Navigational pages</td>
<td>153</td>
</tr>
</tbody>
</table>

**Content pages**
- Article/scientific work: 27
- How-to/tutorial: 25
- Blog: 21
- Other content: 21
- Product: 19
- Announcement: 18
- Info data: 15
- Info text: 14
- About: 11
- Manual guide: 9
- FAQs/Q&A: 8
- Search results: 7
- Teaching materials: 7
- Contact: 4
- Review: 4
- Encyclopaedia entry: 3
- Info advice: 3
- Visual/multimedia: 3
- Total: 219

**Transactional pages**
- Download: 25
- Log-in: 23
- Search interface: 16
- Maps: 13
- Library catalogue: 9
- Translator: 5
- Dictionary: 3
- Calculator: 2
- Other: 2
- Registration form: 1
- Test: 1
- Total: 100

**Main pages**
- Organisation homepage: 101
- Corporate homepage: 84
- Portals and other: 78
- News and serials: 41
- Personal homepage: 23
- E-shop: 19
- Total: 346

**Navigational pages**
- Index: 67
- Table of contents: 25
- Product catalogue: 23
- Resource page: 21
- Mailing list archive: 7
- Archive index: 4
- Directory: 3
- Repository: 3
- Total: 153
A final observation can be made both on relevance and research into web genres. From a system point of view, relevance has been seen as a matter of topicality or best match between the user query and the document subject matter, understanding topicality as “something a document has” more than “something ascribed to documents by persons with some qualifications from some perspective and for some purpose” (Hjørland, 2010, p. 227). Research into web genres seems to approach retrieval with the same assumption, looking for features within documents that match certain users’ expectations, rather than looking at users making actual use of web documents. In this sense research into web genres should probably link up with the research work done after Broder (2002) and Rose and Levinson (2004), specifying the connections between intentions behind queries and web genres. We understand that different query intentions are likely to be answered by different web genres, as genres represent purposive communicative actions. Some of the studies conducted after Broder (2002) and Rose and Levinson (2004) can be especially significant for web genre research. For example Lee et al. (2005) explain that for navigational queries users are likely to visit only one website, whilst for informational queries they will click on several results. If information queries are answered by content pages, as we assumed when analysing the data of this research, then they will be more likely to be used in combination with each other. In contrast web genres answering navigational queries will be more likely to be used in isolation. In this way research into web genres could better capture the behavioural implications of web document use.

We undertook the research to find out if a generic pattern is recognisable in bookmarked webpages. With all the limitations of the sample studied here, we can say that the most notable tendency is for users to bookmark main pages, either homepages or other types of pages that grant access to a website. Homepages can be regarded as multi-functional genres, which do not work for one or a few specific purposes other than generically accessing a website and the services or information it provides. Vaughan et al. (2007), studying links between academic sites, found that links to homepages were less meaningful than other page types in describing the semantic relationship between sites. They discovered that links to non-homepages between Canadian universities’ websites tended to show teaching and research relationships, whereas links to the homepage showed more general relationships between institutions. Despite commonly accepted assumptions in research into automatic recognition of genre, web users probably do not always look for specific webpages when searching the web, but may expect an interaction process supported by more general pages, during which they shape their research question more and more precisely. In this interaction process more general webpages such as homepages, providing plenty of browsing and linking options, play just as important a role as webpages directly addressing a problem. Montesi and Navarrete (2008), studying the web genres used by a software engineer at various stages of his work, observed that at different stages he used different web genres, though he was making relevance decisions at all stages. The nature of the relevance decisions made at each stage was different and dependent on the task at hand. Saracevic (2007a, p. 2130) states that “different selections are made in different stages using similar criteria, but possibly with different weight”. The multi-functionality and elusive nature of many web genres is often considered an impediment in web genre research (Santini, 2007), as it clashes with the need to develop systems able to automatically and unequivocally recognise web genres. However in our opinion, it might be a crucial aspect of some web genres to match users’ expectations with IR systems as multi-functionality might be related to “objective” aspects such as hyperlinks or coexistence of different components, such as
log-in interfaces, text, site maps, etc. Finally, the popularity of homepages may also be explained by their ability to help users assess the recency and authority of information. Taylor et al. (2009) studied the evolution of relevance criteria adopted by users along a search process and found out that, more than any other, the importance of recency, institution, and authority increase along the search process.

Among the four super-genre categories, content pages were the largest group after main pages, accounting for 26 per cent of all classified pages, and the most varied results, with approximately 20 different web genres classified. This leads to two interrelated observations. On the one hand, it explains the tendency of the literature to study single genres within this category, such as blogs (Herring et al., 2004, 2005), reviews (Pollach, 2006), FAQs/Q&A (Cox and Morris, 2004), or wikis and online encyclopedias (Emigh and Herring, 2005). On the other hand, it uncovers a gap in the literature that has disregarded other web genres, especially those classified at a transactional and navigational super-genre level. In this research, transactional pages appeared to be bookmarked for re-use purposes more than any others according to the participants’ interviews. What we know (Hendry et al., 2006) about resource pages is one of the few contributions on web genres other than those listed under content pages or main pages. The web genres included under “main pages” have been widely studied as well, especially homepages, being considered the only truly web genre with no offline antecedents. Navigational web genres probably deserve more attention from the research community as genres supporting browsing and linking combinations of web genres, such as homepages. Studying information seeking on the web, Kari (2004) complains that, despite the fact that browsing has been found repeatedly to be a more common activity on the web than searching, most of the research has concentrated on querying. Furthermore, browsing appears to be based on different relevance criteria than searching. Bodoff (2006) speaks of two types of relevance: relevance for browsing and relevance for searching. In his research, he found a negative correlation between the two types of relevance, as search relevance depends strongly on topicality whereas browsing relevance depends strongly on non-topical factors. Studying web genres that support browsing would make a contribution to this gap in the literature, and could improve web genre retrieval.

**Conclusion**

Analysing webpages in terms of genre is a very complex task. Web genres are numerous and diverse, and it is often difficult to assign them a significant name likely to be universally recognised and accepted. The purpose of the research was to find out if a generic pattern is recognisable in the webpages bookmarked during interaction with the web. The results obtained from the analysis of 833 web links and interviews with 17 participants show that a classification at a super-genre level is a more viable strategy with which to look at webpages as genres, and that users tend to bookmark main or access pages more often than content, navigational, or transactional pages. That users might be interested in more general pages such as main pages should be taken into account in future research into web genres, which has so far disregarded the fact that web users often expect to interact with the web through browsing and not always or exclusively through searching. Further research should also be carried out on web genres different from content pages, such as transactional or navigational pages. Finally, we consider that linking up with research into intentions behind queries would considerably improve experimentation into retrieval systems based on the concept of genre, as it would allow researchers to more closely capture the behavioural implications.
of search and web document use. The results also cast some light on the concept of relevance, which appears to be a time-bounded concept, strongly related to use or utility, and happening in real life. It also appears to vary depending on whether the user is searching or browsing. All this means that different relevance criteria might apply for different web genres, depending on intended use and time, and the main activity supported.

References


Bakhtin, M.M. (1986), Speech Genres and Other Late Essays, University of Texas Press, Austin, TX.


About the author
Michela Montesi has been an Assistant Professor in the Complutense University of Madrid’s Faculty of Information Science since January 2008. Previously she conducted post-doctoral research at the University of Amsterdam (2005-2007). Her research fields are scientific communication, information behaviour (especially on the web), reading behaviour, the social sciences and the humanities, and genre theory. Michela Montesi can be contacted at: mmontesi@pdi.ucm.es

To purchase reprints of this article please e-mail: reprints@emeraldinsight.com
Or visit our web site for further details: www.emeraldinsight.com/reprints