CROSS-REFERENCING METHOD AND SYSTEM FOR ONLINE COMMENTARY

For use in a blogging system, there is provided a method of presenting a user with a list of one or more blog entries that are related to a selected blog entry. The method includes populating (300) a database (114) with a corpus of blog entries retrieved from a plurality of blogs. A selected blog entry of the user is identified (202) and a list of blog entries relevant to the selected blog entry is generated. In particular, the list is generated by cross-referencing (204) the selected blog entry with entries in the corpus, and each entry in the list is assigned a corresponding relevance value based upon the cross-referencing. The user is then presented (206) with a list of one or more related blog entries selected from the relevant blog entries in the corpus on the basis of the relevance values. Methods in accordance with the invention are able to cross-reference blog entries automatically, without requiring additional effort on the part of the user. Furthermore, the database may be populated with blog entries received from blogs across the entire blogosphere, and is not therefore limited to cross-referencing blogs hosted, for example, at a single site. The selected blog entry is utilised to identify other blogs and blog entries of potential interest to the user, without requiring the user to conduct an explicit search, or to identify and enter keywords for the purposes of conducting such a search. The invention also provides computer-implemented apparatus (102), and computer software products, for providing a user with a list of blog entries in accordance with the inventive method.
CROSS-REFERENCING METHOD AND SYSTEM FOR ONLINE
COMMENTARY
FIELD OF THE INVENTION

The present invention relates generally to online commentary facilities, commonly known as web logs or "blogs". More particularly, the invention encompasses methods, apparatus and software products for providing an improved service for identifying blogs and blog entries of interest to users.

BACKGROUND OF INVENTION

A web log, more commonly now known simply as a "blog", can generally be characterised as a publicly-accessible online personal journal written by an individual, often referred to as a "blogger". Blogs may be updated daily, or more or less frequently, and an active blog is generally updated on a reasonably regular basis.

A typical blog includes information, identified more generally herein as "content", which relates to one or more specific topics. Many blogs are personal journals or diaries, and as such the "topic" encompasses the daily activities, thoughts, interests, views, opinions, and so forth, of the blogger. Other blogs focus on a particular subject, such as politics, travel, law, technology etc, and may be characterised according to their genre. Many blogs are the work of private individuals, maintained for personal purposes, however other blogs may serve public, political or business purposes.

The popularity of blogs and blogging has grown enormously in recent years, such that there are now tens of millions of blogs published on the global Internet, each of which may contain many thousands of individual entries. This huge collective body of blog content is sometimes referred to as the "blogosphere".

Technological developments that have contributed to the rising popularity of blogging include software for enabling bloggers to set up and maintain their blogs, and software and network-based protocols supporting the publishing and dissemination of information relating to established blogs. Dedicated software applications eliminate or reduce the need for individual bloggers to possess detailed technical knowledge in order to operate a blog. Web-based systems enable bloggers to operate from anywhere on the Internet, and such software
also allows users to create and maintain blogs posted on web servers operated by third-party service providers. Maintenance tools and hosting services are provided by various operators, including web hosting companies, Internet service providers, and Internet portals such as Google.

Additionally, RSS (Really Simple Syndication) publishing features have been incorporated into various blogging applications. The RSS technology allows service providers and individual users to be updated with RSS feeds from selected blogs, so that updates to the blogs are identified, and the most recently published content on any selected blog may be retrieved. It should be noted that alternative technologies to RSS have been developed, providing substantially similar functionality, such as the Atom technology. Newly-created blogs and blog entries may also be identified using alternative internet-based technologies, such as "ping" services, web crawlers and the like.

Accordingly, the creation, maintenance and publishing of blogs is now easily within the reach of most individuals with access to the Internet, and no longer requires significant technical expertise. However, the resulting massive growth in the number of blogs in the blogosphere has created new problems for bloggers and readers of blogs. In particular, the massive volume of blog content now existing has made it increasingly difficult for readers to identify blogs, or individual entries within blogs, containing content of interest. Similarly, the problem facing individual bloggers is that of how to ensure that their blogs can be identified by potential interested readers.

In order to address the aforementioned problems, a number of techniques and technologies have been developed. For example, a number of blog-specific search engines have been created, which allow readers to search for blogs dealing with a selected topic. Search systems specifically developed for searching blogs typically incorporate specialised features, such as the ability to search for content by a specific author, content published on a specific date, or within a specified range of dates, content published under a particular title, and so forth. Otherwise, however, blog search engines suffer from similar disadvantages to other types of search engine, such as general web searching systems, namely the difficulties associated with selecting appropriate keywords for identifying content of interest. As with any search, even if reasonably effective keywords are
entered, the results may include a large number of blogs, or blog entries, that are of little interest to the searcher. Less effective keyword choice may result in even the top search results being of minimal interest.

Accordingly, other approaches for promoting blogs and blog content have been developed, which are broadly based upon the general principle of "cross-references" between blogs. Such approaches are based upon the observation that a reader of one blog is likely to be interested in reading other blogs, covering either similar or dissimilar topics of interest. Accordingly, blogs may be promoted, and drawn to the attention of potential readers, by including corresponding information and links within other blogs.

One relatively unsophisticated cross-referencing system for promoting blogs, and thereby generating blog traffic, is based upon the substantially random inclusion of crosslinks between blogs. For example, a particular blog hosting site may have a large number of subscribing bloggers, each of which would like to attract interest in the content that they are providing via their blogs. The frequency with which a particular subscriber's blog may be presented to readers of other blogs may depend upon factors such as the level of subscription fees paid by the subscriber, and/or the extent to which the subscriber themselves utilises the cross-referencing system to view the blogs of other subscribers. In any case, the general objective is to provide a mechanism whereby subscribers can be assured that their blogs are being presented to readers and other subscribers for consideration and potential viewing.

However, the effectiveness of the substantially random presentation of blogs to potential readers is somewhat limited. In particular, since the likelihood that a randomly selected blog will be of interest to a particular reader is relatively low, the amount of interest and traffic generated by such systems may be minimal. In the worst case, the lack of relevance of most cross-referenced blogs will cause readers to ignore the links presented to them, on the basis of past experience that they are unlikely to be of great interest.

An alternative cross-referencing method is known as "blog rolling". A blog roll is essentially a list of links made by one blogger to the blogs of other, typically like-minded bloggers. It may often be reasonably presumed that a reader who is interested in the content provided by one particular blogger may also be
interested in at least some of the content of interest to that blogger. The links may be reciprocal, in order to provide corresponding benefits to both bloggers associated by a link, and blog rolls may provide "chains" of links potentially directing readers to a wide range of blogs relating to similar or dissimilar topics to that of the original blog.

The main problem with blog rolling is that the process of identifying and recording the blogs of like-minded bloggers may be slow and tedious, and therefore may consume an excessive amount of time that a blogger may prefer to devote to the generation and publication of new content. In order to simplify the task of recording a blog within a blog roll, automated or semi-automated systems such as that provided by www.blogrolling.com have been developed. Such systems maintain blog rolls on behalf of subscribers within a database, and provide relatively simple mechanisms for including the blog roll within the subscriber's blog page. However, such systems do not provide any assistance in identifying blogs suitable for inclusion in the subscriber's blog roll.

Yet another prior art method for cross-referencing blogs attempts to present readers with a list of "related posts" and/or "related blogs" published by other bloggers. The existing method requires each participating subscriber (/e blogger) to provide a list of one or more keywords or "tags" to identify the subject matter of each blog entry created. When a reader views one such "tagged" entry in a particular blog, the system presents a list of related entries from other bloggers, being entries that have the same or similar tags. While this is a potentially useful technique, it suffers from at least two significant problems. Firstly, it relies upon all bloggers providing tags along with their content, and the systems developed to date have made this functionality available only to bloggers at a single hosting site. Accordingly, it is not presently possible to deploy such a method across the entire blogosphere. Furthermore, the assumption that bloggers selecting the same or similar tag implies that the corresponding content is the same or similar may be false. The thought processes employed by two separate bloggers in selecting appropriate tags may be quite dissimilar, resulting in the choice of similar tags in association with different content. Additionally, tag-based systems are open to abuse, insofar as bloggers may deliberately provide tags that are unrelated to the content of a blog entry, but which are associated
with popular content, as a means of artificially elevating the entry within lists of popular topics.

Clearly, there remains a need for an improved method, system and/or apparatus for cross-referencing blogs and blog content, to assist readers in identifying blogs of interest, and to assist bloggers in attracting a readership, which mitigates at least some of the aforementioned problems of the prior art.

SUMMARY OF THE INVENTION

In one aspect, the present invention provides a method in a blogging system of presenting a user with a list of one or more related blog entries relative to a selected blog entry, the method including the steps of:

- populating a database with a corpus of blog entries retrieved from a plurality of blogs;
- identifying a selected blog entry of the user;
- generating a list of blog entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with entries in the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and
- presenting the user with a list of one or more related blog entries selected from the relevant blog entries in the corpus on the basis of the relevance values.

Advantageously, the invention is able to cross-reference blog entries automatically, without requiring any additional effort on the part of a user, ie a blogger or a blog reader. Furthermore, the database may be populated with blog entries retrieved from blogs across the entire blogosphere, and is not therefore limited to cross-referencing blogs hosted, for example, at a single site. Generally, the process of cross-referencing content in different blog entries may be understood as a form of search, wherein a common occurrence of terms in different entries may be utilised to assess a corresponding degree of similarity in the respective content. Accordingly, the selected blog entry of the user is utilised to identify other blogs and blog entries of potential interest, without requiring the user to conduct an explicit search, or to identify and enter keywords for the purposes of conducting such a search. Accordingly, the invention may significantly simplify the task of identifying blogs and blog entries of interest to a
user, while simultaneously substantially reducing the effort required on the part of bloggers and blog readers.

The step of populating the database preferably includes accessing services which provide updated lists of blogs, and notification of changes to blog content. For example, one suitable service is presently available from Weblogs.com. Alternatively, or additionally, new blogs may be added to the database by direct notification or signup by blog owners. Furthermore, if the selected blog entry is not present in the database, then it may be added to the database, along with further content of the blog from which it has been selected.

The step of populating the database may further include the ongoing periodic reading of feeds, such as RSS or Atom feeds, of known blogs listed in the database in order to identify and fetch new entries for inclusion. In general, the step of populating the database should be understood as ongoing, in order to keep the database up-to-date as new blog entries, and new blogs, are added to the blogosphere.

In preferred embodiments, the selected blog entry is an entry selected for viewing by the user, and served to a web browser of the user by a web server hosting the corresponding blog. Accordingly, the step of identifying a selected blog entry preferably includes embedding code in the served page, said code being executable by the user's web browser to notify a blog entry server of the selected blog entry. The blog entry server is then able to retrieve a copy of the selected blog entry, in order to generate a list of blog entries relevant to the selected entry from within the corpus.

In preferred embodiments, the step of generating a list of blog entries by cross referencing includes accumulating co-occurrence values corresponding with common occurrences of stemmed words within the selected blog entry and within other blog entries in the corpus. In this context, it will be understood that the term "stemmed words" refers to words having a common stem, such that, for example, the words "read", "reads", "reading", and "readings" are considered to be equivalent.

Preferably, each co-occurrence value is weighted in accordance with a frequency of occurrence of the corresponding stemmed word within the corpus, such that, for example, a higher co-occurrence value is associated with words
occurring with lower frequency within the corpus. Advantageously, this approach accords greater significance to less common terms within the search query, while affording relatively little significance to common words such as "and", "the", "a" and so forth. Indeed, in preferred embodiments of the invention all words appearing in all documents within the corpus are assigned a specific weighting based upon their frequency of occurrence, thereby avoiding the making of artificial distinctions between predetermined "common" and "uncommon" words. Advantageously, therefore, the selected blog entry effectively acts as a form of "search query", wherein less common words appearing within the selected entry have the greatest significance as "keywords" in the search. However, unlike a conventional search which utilises user-supplied keywords, and generally accords equal significance to each, the preferred cross-referencing method of the present invention may utilise all words appearing in the selected entry, and will automatically determine which of those words are most significant in determining the relevance of other blog entries within the corpus. Each entry in the generated list of blog entries relevant to the selected blog entry is assigned a relevant value, which is preferably determined by accumulating corresponding co-occurrence values weighted in accordance with the probability of occurrence of the respective stemmed words.

The step of presenting the user with a list of one or more related blog entries preferably includes the blog entry server serving back a list which is embedded in a blog page displayed by the user's browser. The embedded list preferably includes clickable links to the corresponding blog entries. The list served back to the browser and embedded in the displayed blog page preferably includes a small number of entries, for example between one and ten entries, having the highest relevance to the selectged blog entry in accordance with the assigned relevance values. Alternatively or additionally the blog entry server may serve a separate web page including a list of relevant blog entries, which may include more entries than the embedded list. In a particularly preferred embodiment, a short embedded list is provided along with a link for selection by the user which opens a separate page including an extended list of relevant entries.
In a system including a plurality of server computers hosting a plurality of blogs, and at least one client device enabling access by a user to said blogs, wherein the server computers and client device are interconnected by a data network, the invention provides, in another aspect, an apparatus for providing a list of one or more blog entries contained within the plurality of blogs, the apparatus including:

- at least one processor;
- a database populated with a corpus of blog entries retrieved from said plurality of blogs;
- at least one network interface connecting the processor to the data network; and
- at least one storage medium containing program instructions for execution by the processor, said program instructions causing the processor to execute the steps of:

  1. receiving information identifying a selected blog entry from the client device via the data network; generating a list of blog entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with entries within the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and
  2. transmitting a list of one or more blog entries to the client device via the data network, said list including entries selected from the relevant blog entries in the corpus on the basis of the relevance values.

The client device typically may be a computer executing web browser software. However, in general suitable client devices include wired or wireless devices connected to the network using various technologies and bandwidths, such as: PC's with wired (eg LAN, cable, ADSL, dial-up) or wireless (eg WLAN, cellular) connections; and wireless portable/handheld devices such as PDA's or mobile/cellular telephones.

In preferred embodiments, the storage medium contains further program instructions causing the processor to execute the step of populating the database with blog entries retrieved from the plurality of blogs. For example, the program instructions may cause the processor to access services via the data network.
which provide updated lists of blogs, and notification of changes to blog content, whereupon the program instructions cause the processor to execute the step of retrieving new and updated blog content and incorporating said content into the corpus.

Transmitting the list of one or more blog entries to the client device may include transmitting the list in a form appropriate for embedding within a blog web page displayed by a web browser on the client device, and/or transmitting the list in the form of a separate web page for display by the web browser.

In a system including a plurality of server computers hosting a plurality of blogs, and at least one client device enabling access by a user to said blogs, wherein the server computers and client device are interconnected by a data network, the present invention provides, in yet another aspect, an apparatus for providing a list of one or more blog entries contained within the plurality of blogs, the apparatus including:

means for receiving information identifying a selected blog entry from the client device via the data network;
means for generating a list of entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with entries within the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and
means for transmitting a list of one or more blog entries to the client device via the data network, said list including entries selected from the relevant blog entries in the corpus on the basis of the relevance values.

The apparatus preferably includes a server computer, connected to the data network, which has at least one processor, and is configured to communicate with client devices, such as computers, and other server computers via the network. The server computer also preferably includes and/or is associated with a database containing the corpus of blog entries.

The means for receiving information identifying a selected blog entry, and the means for transmitting a list of one or more blog entries, may include suitable network interface hardware of the server for interfacing to the data network, and may further include one or more software components executed by the processor, the software components including instructions to effect the receiving
of information identifying the selected blog entry, and the transmitting of the list of one or more blog entries.

The means for generating a list of blog entries relevant to the selected blog entry typically includes one or more software components executed by the processor.

In particularly preferred embodiments, the apparatus further includes means for populating the database with the corpus of blog entries, for example by accessing services via the data network which provide updated lists of blogs, and notifications of changes to blog content.

In yet another aspect, the present invention provides a computer software product including a computer-readable medium bearing program instructions for causing at least one processor to execute a method, in a blogging system, of presenting a user with a list of one or more related blog entries relevant to a selected blog entry, the method including the steps of:

- populating a database with a corpus of blog entries retrieved from a plurality of blogs;
- identifying a selected blog entry of the user;
- generating a list of blog entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with entries in the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and
- presenting the blog user with a list of one or more related blog entries selected from the relevant blog entries in the corpus on the basis of the relevance values.

Furthermore, as will be appreciated by those skilled in the art, and particularly by those involved in the blogging community, individual bloggers often have a desire to identify other like-minded bloggers within the blogosphere. Accordingly, in yet another aspect, the present invention provides a method in a blogging system for enabling a first blogger to identify one or more further bloggers and/or blogs of possible interest, the method including the steps of:

- populating a database with a corpus of blog entries retrieved from a plurality of blogs;
identifying a selected blog entry created by the first blogger which relates to a topic of interest of the first blogger;

generating a list of relevant blog entries relating to said topic of interest from within the corpus by cross-referencing the selected blog entry with entries in the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and

presenting the first blogger with at least one results list including one or more of:

- a list of blog entries created by one or more further bloggers, said entries being selected from the relevant blog entries on the basis of the relevance values;

- a list of one or more blogs maintained by one or more further bloggers, said blogs being identified from corresponding blog entries selected from the relevant blog entries on the basis of the relevance values; and/or

- a list of one or more further bloggers, said bloggers being identified from corresponding blog entries selected from the relevant blog entries on the basis of the relevance values.

Further preferred features and advantages of the invention will be apparent to those skilled in the art from the following description of preferred embodiments of the invention, which should not be considered to be limiting of the scope of the invention as defined in the preceding statements, or in the claims appended hereto.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of the invention are described with reference to the accompanying drawings, in which like reference numerals refer to like features, and wherein:

Figure 1 illustrates an exemplary blogging system embodying the present invention;

Figure 2 is a flowchart illustrating a preferred method of presenting a user with blog entries in accordance with the invention;

Figure 3 is a flowchart illustrating a preferred method of populating a database with a corpus of blog entries in accordance with the present invention;
Figure 4 illustrates an exemplary blog entry page incorporating an embedded list of related blog entries according to an embodiment of the present invention; and

Figure 5 is a sample script template for inclusion in a blog page template, according to an embodiment of the present invention.

**DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS**

Figure 1 illustrates an exemplary blogging system 100 embodying the present invention. The system 100 includes a blog entry server 102 which is configured, as described in greater detail below, in accordance with an embodiment of the present invention to present a blog user with a list of one or more related blog entries which are relevant to a selected blog entry. The blog entry server 102 is connected via the Internet 106 to a plurality of other client and server systems, including a blogger client computer 104, a blog hosting server 108, a blog reader client computer 110, and an update service server, alternatively known as a "ping" server, 120.

It will be appreciated that Figure 1 is exemplary only and depicts the blogging system 100 schematically. In describing this exemplary embodiment, it is not intended to limit the technology employed in the servers, clients and/or communications links. For example, the client computers are representative of a range of possible user terminals, which may be wired or wireless devices connected to the network using various technologies and bandwidths. For example, applicable user terminals include (without limitation): PC's with wired (eg LAN, cable, ADSL, dial-up) or wireless (eg WLAN, cellular) connections; and wireless portable/handheld devices such as PDA's or mobile/cellular telephones. The protocols and interfaces between the user terminals and the servers may also vary according to available technologies, and include (again without limitation): wired TCP/IP (Internet) protocols; GPRS, WAP and/or 3G protocols (for handheld/cellular devices); Short Message Service (SMS) messaging for digital mobile/cellular devices; and/or proprietary communications protocols.

The interactions amongst the various client and server computers within the exemplary system 100 will now be described.

For the purposes of illustrating the invention, Figure 1 includes a single blogger client computer 104, upon which a web browser software application is
executing. A server 108 provides blog hosting services to a user of client computer 104, who owns and operates at least one blog hosted by the server 108. Hereafter, this exemplary blogging user is referred to simply as the blogger, and the server 108 as the blog host. The blogger maintains a blog on the blog host 108 via a provided web interface utilising the web browser software executing on the blogger client computer 104. A number of web-based blog hosting services are now available via the Internet, and will be familiar to persons skilled in the art. Accordingly, it is unnecessary to provide further details of the web-based blog hosting service herein.

Client computer 110 also executes web browser software, and is operated by a blog reader. The reader is able to access various blogs in the blogosphere, including the blog hosted by server 108 that is owned and maintained by the blogger using client computer 104. For the purposes of illustrating the present invention, it is assumed that the user of client computer 110 is a reader of this blog, and is generally desirous of identifying further blogs and blog entries covering similar or related topics. In this regard, it will be appreciated that there are, connected to the Internet 106, numerous further blog hosting systems, similar to the server 108, collectively supporting some tens of millions of individual blogs, each consisting of multiple blog entries, owned and published by millions of bloggers similar to the user of client computer 104. Accordingly, it is a non-trivial problem for the blog reader using client computer 110 to identify within this massive blogosphere additional blogs and blog entries that may be of interest. Similarly, bloggers such as the user of client computer 104 face a significant problem of gaining sufficient visibility and/or publicity for their blogs in order to attract interested readers. The blog entry server 102 is therefore provided to assist both reader and blogger address their respective problems in this regard.

To this end, the blog entry server 102 incorporates a specialised form of search engine for identifying relevant blog entries, to enable lists of related entries to be incorporated within the blog pages of bloggers, and presented to blog readers. The server 102 includes at least one processor 112, as well as a database 114, which would typically be stored on a secondary storage device of the server 102, such as one or more hard disk drives. The database 114 is populated with a corpus of blog entries retrieved from a plurality of blogs
distributed throughout the blogosphere. Suitable methods for populating the
database are described in greater detail below, with reference to Figure 3.

Blog entry server 102 further includes at least one storage medium 116, typically being a suitable type of memory, such as random access memory, for containing program instructions and transient data related to the operation of the blog entry searching service, as well as other necessary functions of the server 102. In particular, the memory 116 contains a body of program instructions 118 implementing a method of identifying and presenting blog entries of interest to a blog reader, in accordance with embodiments of the invention. Additionally, the body of program instructions 118 includes instructions for providing a web-based interface to the blog entry service, and a cross-reference engine, the operation of which will be described hereafter, for identifying blog entries of interest to a blog reader within the corpus of entries stored in database 114.

Also illustrated in Figure 1, and connected via the Internet 106 to the exemplary system 100, is a ping server 120, which may be used to assist in populating and maintaining the database 114, as will be described in greater detail with reference to Figure 3.

Figure 2 shows a flowchart 200 which illustrates a preferred method of presenting a blog reader with a list of one or more blog entries of interest in accordance with the present invention. In particular, the method is based upon a blog reader, such as the user of client computer 110, selecting a particular blog and blog entry, such as those published by the user of client computer 104 via blog host server 108, for viewing. It is assumed that this selected blog entry is of interest to the blog reader, and the goal of the method represented by the flowchart 200 is to identify and present to the reader a list of further blogs and/or blog entries that may also be of interest, by virtue of a similarity in subject matter.

In order to achieve the abovementioned outcome, at step 202 details of the selected blog entry are received, via the Internet 106, by the blog entry server 102. This may readily be achieved, as described in greater detail below with reference to Figure 5, by including appropriate script code within the blog entry page served to the browser on the client computer 110. Execution of the
script code by the browser results in the delivery of details of the selected blog entry to the blog entry server 102.

Once the blog entry server 102 has received details of the selected blog entry, it is able to retrieve a copy of the content of the selected blog entry itself. The blog entry content may be retrieved from the blog hosting server 108, however in many cases a copy will already be contained within the corpus of blog entries held in the database 114. In either case, the full text of the blog entry is employed by the blog entry server 102, in accordance with step 204 of the method illustrated by flowchart 200, in order to rank other relevant blog entries included within the database 114 using a cross-reference engine. In particular, in accordance with an exemplary embodiment, the cross-reference engine uses the content of the selected blog entry as a "central" document which is cross-referenced with other documents (ze blog entries) held within the corpus contained in database 114. More specifically, a relevance value may be computed for each blog entry in the corpus relative to the content of the selected blog entry, by accumulating co-occurrence values corresponding with common occurrences of stemmed words within the selected blog entry and within each further blog entry within the corpus. Generally speaking, therefore, entries including a greater number of stemmed words in common with the selected blog entry may be more highly ranked, and considered to be of greater interest than entries including fewer stemmed words in common with the selected blog entry.

In itself, however, this simple approach is not sufficient to identify blog entries of genuine interest, because it fails to distinguish between those words within the selected blog entry that are of particular significance to the reader's interests, and those that are largely irrelevant, including such frequently occurring words as "and", "the", "a" and so forth.

Accordingly, the preferred implementation of the cross-reference engine applies a weighting to each occurrence value in accordance with the frequency of occurrence of the corresponding stemmed word within the corpus. Specifically, every stemmed word within the corpus is assigned a weighting value based upon the probability of occurrence of the stemmed word, wherein the probability of occurrence p is defined as being the number of documents containing at least one occurrence of the stemmed word divided by the total number of documents in
the corpus. As will be appreciated, common words such as "the" will have a probability of occurrence substantially equal to one. Conversely, highly uncommon words may have an extremely low probability of occurrence, particularly if the corpus is very large, as it may be when populated with a significant proportion of all blogs and blog entries available in the blogosphere.

As will be appreciated, words appearing in the selected blog entry that have a relatively low probability of occurrence provide more information regarding the reader's interests than very commonly occurring words. This is because a rare word is more effective in identifying documents that are likely to be of interest in relation to the selected blog entry. By applying an information theoretic approach, a suitable weighting value may be assigned to words appearing within the corpus of blog entries based upon the probability of occurrence of each stemmed word. In particular, according to the preferred embodiment of the cross-reference engine, the inverse of the probability of occurrence of each stemmed word is used as the basis for defining a corresponding number of bits of information notionally associated with the word. To be precise, the weighting value \( w \) of a stemmed word having a probability of occurrence \( p \) is calculated according to the following formula:

\[
w = \log_2(p^{-1})
\]

By accumulating the co-occurrence values computed by the cross reference engine in this manner, every word in the corpus at any given time may be assigned a corresponding ranking value in relation to the selected blog entry. Blog entries, and the blogs from which they are extracted, having a higher ranking value are thereby considered to be of greater interest to the reader than documents having a lower ranking value. The ranking values may therefore be used as corresponding relevance values which may, in principle, be assigned to any or all of the blog entries contained within the corpus held in database 114.

At step 206 of the method illustrated by flowchart 200, the blog entry server 102 serves a list of high-ranking related blog entries to the reader via the Internet 106. In accordance with a particularly preferred embodiment, an initial short list of high-ranking results is embedded within the blog entry page displayed by the web browser on the client computer 110, under the control of the
aforementioned script code included within the page served by the blog host server 108. Further details of this mechanism are provided below with reference to Figure 5. Although a simple, and perfectly practical, approach would be to provide the reader with a short list including the most highly ranked blog entries in order, it may be preferable to employ an alternative approach. In particular, the blog entry server 102 may provide a short list of, say, two or three blog entries selected at random out of, for example, the top fifty most highly ranked and most recently published blog entries. This approach ensures that the list of related entries is relatively current, and that the reader will be presented with a slightly different short list of related blog entries each time the selected blog entry or another blog entry relating to the same or similar topic within the same or a different blog, is viewed. The value in providing this greater level of variety is that it increases the opportunity for the reader to identify a variety of potentially relevant and interesting blogs, and similarly increases the potential exposure of these blogs for the benefit of their corresponding bloggers.

While the foregoing discussion with reference to Figure 2 describes a method of presenting a blog reader with a list of one or more blog entries of interest based upon a selected blog entry, it will be appreciated that in various embodiments the invention is equally applicable to bloggers. Indeed, it is reasonable to presume that most, if not all, bloggers also read the blogs of other bloggers. In particular, those familiar with the blogging community will understand that a common desire of bloggers is to identify like-minded individuals within that community.

In order to satisfy this need, and to assist a first blogger to identify one or more bloggers and/or blogs of possible interest, the selected blog entry used as the basis for the cross-referencing process may be an entry created by the first blogger which, it may be assumed, relates to a topic of interest to the blogger. (It will be appreciated that the entry created by the blogger may be either complete or incomplete, ie that the selected entry may be one that the blogger is in the process of composing.) By employing the same cross-referencing method previously described, a list of relevant blog entries relating to the topic of interest may be generated from within the corpus.
Depending upon the particular requirements of the blogger, the results of this process may be presented in the form of a list of blog entries created by one or more further bloggers, a list of references to the corresponding blogs and/or a list of the corresponding bloggers responsible for maintaining those blogs. The first blogger may then choose, for example, to read one or more of the individual identified blog entries, to review the corresponding blogs, to seek out additional blogs or other content produced by the respective further blogger(s), and/or to contact the bloggers e.g. via email, if relevant contact details are available. Alternatively or additionally, the first blogger may choose to crosslink to one or more of the identified blogs, for example via a blogrolling mechanism such as that described later with reference to Figure 4.

Figure 3 shows a flowchart 300 which illustrates a preferred method of populating the database 114 with the corpus of blog entries. At step 302, the blog entry database 114 is established, and initially the corpus may be empty. At step 304 new blogs and/or blog entries are identified by the blog entry server 102. At step 306, the new entries are retrieved, and the corpus is updated to include details of the newly identified blogs and blog entries. As will be appreciated, the execution of step 306 is relatively straightforward, once a new blog and/or new blog entry has been identified. A number of mechanisms may be employed to perform this identification within step 304.

A first, and generally preferred method of identifying new blogs and blog entries utilises a third party update service. An update, or ping server, 120, which provides such a service, is included within the exemplary system 100 illustrated in Figure 1. Such services are already operating via the Internet 106, with just one example being the Weblogs.com service provided by VeriSign. Blog hosting servers, such as the server 108, are readily configured to transmit messages, commonly known as "pings", to the ping server 120 to notify it of the addition and publication of new content, such as the creation of new blogs and new blog entries. The ping server 120 then publishes an update file periodically, which may be retrieved by the blog entry server 102. For example, the Weblogs.com service publishes an XML file every five minutes, which lists the URLs of all blogs for which updates were notified during the previous five minutes. According to the preferred embodiment, the blog entry server 102 downloads the update file
every five minutes, and creates an internal list of new and updated blogs that require reading. A separate executing process then works through this list fetching and parsing the RSS or Atom feed of the new and updated blogs, identifying any and all new entries from the feeds, and adding them to the corpus contained within the database 114.

An alternative method that could be employed to maintain the corpus would be to effectively incorporate a ping service within the blog entry server 102. It is envisaged, however, that this would be a less effective solution, since it would require blog hosts, such as the server 108, to be configured to send pings to the blog entry server 102. Since existing update services, such as Weblogs.com, are already well-known and widely utilised, it would appear undesirable to establish a competing service, although it should be understood that such a mechanism is within the scope of the invention.

One further mechanism for identifying new and updated blogs is via the selected blog entries themselves. For example, if the reader using the client computer 110 accesses a blog entry on server 108 that is not already contained within the database 114, the blog entry server 102 is able to conclude that its records of the corresponding blog are not up-to-date. It may then add the blog to the list of new and updated blogs marked as requiring an update read. The records of these blogs within the corpus will then be updated in due course along with other new and updated blogs identified via the ping server 120.

Figure 4 illustrates an exemplary blog entry page 400 that may be served by the blog host server 108 and displayed via the browser software executing on client computer 110. The blog entry page 400 includes centrally the contents of the blog entry itself 402. Also embedded in the page 400, in a right-hand column, is a blog roll 404. The blog roll consists of a list of blogs identified and selected by the blogger responsible for the displayed blog. As such, the reader will understand that each of the blogs listed in the blog roll is of interest to the blogger responsible for publishing the present entry, and may therefore similarly be of interest to the reader. Blog rolls of this type are known in the prior art, however their functionality may be enhanced, as described hereafter, by integration within embodiments of the present invention.
In a left-hand column of the page 400 there is embedded a list of related posts 406. This list has been provided by the blog entry server 102 based upon the content of the selected blog entry 402. As will be seen in Figure 4, only two related entries have been listed, exemplified by the first entry 408 entitled "Where has all the water gone?". The list of related entries is integrated with the blog roll feature by providing a "blogroll it" link 410. The link 410 provides a shortcut to a blog roll management page which enables the user to add the corresponding blog to their blog roll. For example, if the author of the present blog entry were to click on this link, it would facilitate the addition of the corresponding blog to the blog roll list 404.

Finally, the related entries list 406 includes a further link 412 which enables a longer list of related entries to be provided. Specifically, the link 412 results in an HTTP request being sent to the blog entry server 102, which responds by serving a separate web page including a more extensive list of the top-ranked related blog entries. Advantageously, this page may also contain additional information and content, such as banner advertisements. Such advertisements may be served by advertising service providers, such as Google or Yahoo, which attempt to select and embed advertising within a web page that is itself chosen on the basis of the web page content. As will be appreciated, participation in such advertising programs may provide a useful mechanism for raising revenue for operating the blog entry server 102.

Figure 5 is a sample script template listing which may be included in a blog page template in order to embed a related entry listing such as the listing 406 embedded within the web page 400. As will be appreciated by persons skilled in the art, the script template contains code to initialise parameters identifying the blog entry page on the Internet, the dimensions, colour, font, and any other desired characteristics of the related entry list 406, and incorporates script code that would typically be hosted and maintained on the blog entry server 102. As previously indicated, the web browser software executing on the client computer 110 executes the script code embedded in the blog entry page, which transmits details of the blog entry to the blog entry server 102, which responds by serving the actual content, consisting of the list of related blog entries 406, for display within the blog entry page.
As will be appreciated, numerous variations of the invention, as described in particular embodiments herein, are conceivable. For example, it will be understood that the functionality provided by the blog entry server 102 may, in practice, be distributed and/or duplicated over a number of server computers that may be geographically co-located or remotely located from one another. Furthermore, it will be understood that while an exemplary system 100 including a blogger client computer 104 and reader client computer 110 has been described, in practice many users will be both bloggers and blog readers and that there need be no essential difference between the client computer systems 104, 110 and the web browser software executing thereon. Accordingly, the present invention envisages a range of client/server arrangements, such as would be apparent to persons skilled in the relevant art.

In other variations, alternative rating or ranking methods may be employed for determining suitable relevance values of blog entries. For example, one advantageous approach involves associating a thesaurus with the cross-reference engine, such that it is possible to accumulate co-occurrence values not only between matching stemmed words, but also between synonyms thereof. Such an approach would enable the improved identification and ranking of blog entries that may relate to subject matter of interest to a user, but which do not use identical terms to those appearing within the selected blog entry. Similarly, by associating one or more multi-lingual dictionaries with the cross-reference engine it would be possible, at least in principle, to identify and rank relevant blog entries in different languages. In any case, it will of course be appreciated that the invention is not limited to any particular language, whether or not a multi-lingual cross-reference engine is implemented.

A further possible enhancement may involve the automated generation of keywords associated with particular blog entries. Advantageously, this approach could provide for a degree of compatibility with prior art methods of identifying relevant blogs and blog entries based upon the use of "tags", while also mitigating the problems of such methods arising from the inconsistent selection of tags by different users. For example, instead of requiring users to select their own tags or keywords, a method in accordance with the present invention could be employed to identify groups of related blog entries, and then keywords generated based
upon the most commonly co-occurring words within the related entries. Such an approach may be particularly powerful when combined with functionality for identifying and cross-referencing synonyms, such that it would become less important for different bloggers to use the same terms when discussing like concepts.

It will therefore be understood that while an exemplary embodiment of the invention has been described herein, this should not be considered to limit the scope of the invention, as defined by the claims appended hereto.
CLAIMS:

1. A method in a blogging system of presenting a user with a list of one or more related blog entries relative to a selected blog entry, the method including the steps of:
   - populating a database with a corpus of blog entries retrieved from a plurality of blogs;
   - identifying a selected blog entry of the user;
   - generating a list of blog entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with entries in the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and
   - presenting the user with a list of one or more related blog entries selected from the relevant blog entries in the corpus on the basis of the relevance values.

2. In a system including a plurality of server computers hosting a plurality of blogs, and at least one client device enabling access by a user to said blogs, wherein the server computers and client device are interconnected by a data network, an apparatus for providing a list of one or more blog entries contained within the plurality of blogs, the apparatus including:
   - at least one processor;
   - a database populated with a corpus of blog entries retrieved from said plurality of blogs;
   - at least one network interface connecting the processor to the data network; and
   - at least one storage medium containing program instructions for execution by the processor, said program instructions causing the processor to execute the steps of:
     - receiving information identifying a selected blog entry from the client device via the data network;
     - generating a list of blog entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with
entries within the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and

transmitting a list of one or more blog entries to the client device via the data network, said list including entries selected from the relevant blog entries in the corpus on the basis of the relevance values.

3. In a system including a plurality of server computers hosting a plurality of blogs, and at least one client device enabling access by a user to said blogs, wherein the server computers and client device are interconnected by a data network, an apparatus for providing a list of one or more blog entries contained within the plurality of blogs, the apparatus including:

means for receiving information identifying a selected blog entry from the client device via the data network;

means for generating a list of entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with entries within the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and

means for transmitting a list of one or more blog entries to the client device via the data network, said list including entries selected from the relevant blog entries in the corpus on the basis of the relevance values.
4. A computer software product including a computer-readable medium bearing program instructions for causing at least one processor to execute a method, in a blogging system, of presenting a user with a list of one or more related blog entries relevant to a selected blog entry, the method including the steps of:

   populating a database with a corpus of blog entries retrieved from a plurality of blogs;
   identifying a selected blog entry of the user;
   generating a list of blog entries relevant to the selected blog entry from within the corpus by cross-referencing the selected blog entry with entries in the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and
   presenting the blog user with a list of one or more related blog entries selected from the relevant blog entries in the corpus on the basis of the relevance values.

5. A method in a blogging system for enabling a first blogger to identify one or more further bloggers and/or blogs of possible interest, the method including the steps of:

   populating a database with a corpus of blog entries retrieved from a plurality of blogs;
   identifying a selected blog entry created by the first blogger which relates to a topic of interest of the first blogger;
   generating a list of relevant blog entries relating to said topic of interest from within the corpus by cross-referencing the selected blog entry with entries in the corpus, each entry in said list being assigned a corresponding relevance value based upon said cross-referencing; and
   presenting the first blogger with at least one results list including one or more of:

      a list of blog entries created by one or more further bloggers, said entries being selected from the relevant blog entries on the basis of the relevance values;
a list of one or more blogs maintained by one or more further bloggers, said blogs being identified from corresponding blog entries selected from the relevant blog entries on the basis of the relevance values; and/or

a list of one or more further bloggers, said bloggers being identified from corresponding blog entries selected from the relevant blog entries on the basis of the relevance values.
START

Receive selected blog entry details from user

Rank relevant blog entries using cross-reference engine

Serve high-ranking results to user

END

Figure 2
Figure 3
The Politics of Water - Introduction

We live in the driest continent, and hence water is a scarce resource. It should therefore not be treated as a nearly free resource, whether in the cities (private and industrial use); or the country, (private and agricultural use).

Are there any people who disagree with this? I suspect that communities would tend to agree, and individuals to disagree. Individuals think about the potential cost of their own consumption; but communities think about the cost of other people’s consumption.

posted by waterchamp at 3:12 PM 6 comments
<script type="text/javascript">
<!--
blogknots_user="<userid>"; //optional for registered users
blogknots_feed_url="<the URL of this feed>";
blogknots_item_url="<the URL of this item or entry>";
blogknots_gb_width="160";
blogknots_gb_height="310";
blogknots_gb_max_results="3";

//color codes musn't contain the '\' sign in front
blogknots_gb_color_border="004371";
blogknots_gb_color_bg="EBF6E2";
blogknots_gb_color_link="004371";
//put examples of font-sizes and font-faces on our page
blogknots_gb_size_link="11";
blogknots_gb_face_link="Verdana";
//-->
</script>
<script type="text/javascript" 
src="http://www.blogknots.com/blogknots/pages/doKnot.js"></script>

Figure 5
INTERNATIONAL SEARCH REPORT

INTERNATIONAL SEARCH REPORT

A CLASSIFICATION OF SUBJECT MATTER

Int Cl


According to International Patent Classification (IPC) or to both national classification and IPC

B FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPCG-G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

USPTO, DWPI G06F & keywords weblog, blog, article, post, document, diary, match, similar, relevant, rank, rate, index and similar terms

C DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>US 6463430 B1 (BRADY ET AL ) 8 October 2002 – Entire document (see particularly figures 1-12, column 2, line 50 to column 3, line 9, column 5, lines 3 to 67, column 8, lines 22 to 46, column 10, line 12 to column 11, line 4)</td>
<td>1-5</td>
</tr>
<tr>
<td>Y</td>
<td>US 2004/0059708 A1 (DEAN ET AL ) 25 March 2004 – Entire document (see particularly figures 1-5, paragraphs 001 8, 0033, 0042, 0047, 0048, 0052, 0053)</td>
<td>1-5</td>
</tr>
<tr>
<td>Y</td>
<td>WO 2006/128136 A2 (INSIDER PAGES) 30 November 2006 – Entire document (see particularly figures 1-4, page 7, line 32 to page 8, line 11, page 14, lines 13-24)</td>
<td>1-5</td>
</tr>
</tbody>
</table>

Further documents are listed in the continuation of Box C

See patent family annex

Date of the actual completion of the international search

20 February 2008

Date of mailing of the international search report

3 March 2008

Name and mailing address of the ISA/AU

AUSTRALIAN PATENT OFFICE
PO BOX 200, WODEN ACT 2606, AUSTRALIA
E-mail address pct@ipaustralia.gov.au
Facsimile No +61 2 6283 7999

Authorized officer

Benjamin Lam

AUSTRALIAN PATENT OFFICE
(ISO 9001 Quality Certified Service)
Telephone No (02) 6225 6121

Form PCT/ISA/210 (second sheet) (April 2007)
This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>US 6463430</td>
<td>NONE</td>
</tr>
<tr>
<td>US 2004059708</td>
<td>AU 2003272687</td>
</tr>
<tr>
<td></td>
<td>AU 2003275251</td>
</tr>
<tr>
<td></td>
<td>AU 2003276935</td>
</tr>
<tr>
<td></td>
<td>BR 0314719</td>
</tr>
<tr>
<td></td>
<td>BR 0314727</td>
</tr>
<tr>
<td></td>
<td>BR PI03 14723</td>
</tr>
<tr>
<td></td>
<td>CA 2499669</td>
</tr>
<tr>
<td></td>
<td>CA 2499768</td>
</tr>
<tr>
<td></td>
<td>CN 1685337</td>
</tr>
<tr>
<td></td>
<td>CN 1689002</td>
</tr>
<tr>
<td></td>
<td>CN 1701331</td>
</tr>
<tr>
<td></td>
<td>CN 1759388</td>
</tr>
<tr>
<td></td>
<td>EP 1552422</td>
</tr>
<tr>
<td></td>
<td>EP 1552436</td>
</tr>
<tr>
<td></td>
<td>EP 1574037</td>
</tr>
<tr>
<td></td>
<td>KR 2005008641</td>
</tr>
<tr>
<td></td>
<td>KR 2006003557</td>
</tr>
<tr>
<td></td>
<td>US 7136875</td>
</tr>
<tr>
<td></td>
<td>US 2004167928</td>
</tr>
<tr>
<td></td>
<td>US 200429758</td>
</tr>
<tr>
<td></td>
<td>US 2006128136</td>
</tr>
<tr>
<td></td>
<td>US 2006284744</td>
</tr>
<tr>
<td></td>
<td>WO 2004029758</td>
</tr>
<tr>
<td></td>
<td>WO 2004030338</td>
</tr>
<tr>
<td></td>
<td>WO 2006284873</td>
</tr>
</tbody>
</table>

Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.

END OF ANNEX