Research interests: their dynamics, structures and applications in unifying search and reasoning

Zeng, Y.; Zhou, E.; Wang, Y.; Ren, X.; Qin, Y.; Huang, Z.; Zhong, N.

published in
Journal of Intelligent Information Systems
2011

DOI (link to publisher)
10.1007/s10844-010-0144-1

document version
Publisher's PDF, also known as Version of record

J Intell Inf Syst (2011) 37:65–88
DOI 10.1007/s10844-010-0144-1

Research interests: their dynamics, structures and applications in unifying search and reasoning

Yi Zeng · Erzhong Zhou · Yan Wang · Xu Ren · Yulin Qin · Zhisheng Huang · Ning Zhong

Received: 11 April 2010 / Revised: 20 October 2010 / Accepted: 23 November 2010 /
Published online: 8 December 2010

Abstract Most scientific publication information, which may reflects scientists' research interests, is publicly available on the Web. Understanding the characteristics of research interests from previous publications may help to provide better services for scientists in the Web age. In this paper, we introduce some parameters to track the evolution process of research interests, we analyze their structural and dynamic characteristics. According to the observed characteristics of research interests, under the framework of unifying search and reasoning (ReaSearch), we propose interests-based unification of search and reasoning (I-ReaSearch). Under the proposed I-ReaSearch method, we illustrate how research interests can be used to improve literature search on the Web. According to the relationship between an author's own interests and his/her co-authors interests, social group interests are also used to refine the literature search process. Evaluation from both the user satisfaction and the scalability point of view show that the proposed I-ReaSearch method provides a user centered and practical way to problem solving on the Web. The efforts provide some hints and various methods to support personalized search, and can be considered as a step forward user centric knowledge retrieval on the Web. From the standpoint of the Active Media Technology (AMT) on the Wisdom Web, in this paper, the study on the characteristics of research interests is based on complex networks and human

Y. Zeng (✉) · E. Zhou · Y. Wang · X. Ren · Y. Qin
International WIC Institute, Beijing University of Technology,
Beijing, 100124, People's Republic of China
e-mail: yizeng@bjut.edu.cn

Z. Huang
Division of Mathematics and Computer Science, Vrije University Amsterdam,
De Boelelaan 1081, 1081 HV, Amsterdam, the Netherlands
e-mail: huang@cs.vu.nl

N. Zhong
Department of Life Science and Informatics, Maebashi Institute of Technology,
Maebashi-City, 371-0816, Japan
e-mail: zhong@maebashi-it.ac.jp

dynamics, which can be considered as an effort towards utilizing information physics to discover and explain the phenomena related to research interests of scientists. The application of research interests aims at providing scientific researchers best means and best ends in an active way for literature search on the Web.

Keywords Research interest detection · Retained interest · Interest duration · Web search refinement · Unifying search and reasoning

1 Introduction

Scientific researchers form a very large community in the Web age, and various services has been provided for them to support their research on the Web platform (Shneiderman 2008), such as Web-based literature search systems (e.g. Google Scholar, CiteSeerX) and researchers online networks (e.g. ResearchGATE). Many of the systems and platforms are based but lack of deeper analysis on the interests of the researchers from the perspectives of their dynamic and structural characteristics. Understanding the nature and models of research interests from these two perspectives may help to produce better and active services for scientists.

We first introduce some measurement parameters to track and portray the changing process of research interests. Then we investigate on how to obtain retained interests over time through an analogy between cognitive memory retention (Ebbinghaus 1913) and interest retention. As a step forward retained interests based on one's own previous information, we also give a preliminary study on the relationship of a user's retained interests and his/her group retained interests in the collaborative network, which reflect his/her co-author group's retained interests. In addition, we investigate on the structure of research interest in a network setting. By using network theory, we provide some understanding on the statistical characteristics on the structure of research interests. Considering from the time perspective, the appearance and disappearance of research interests is also a dynamic process. Inspired by the quantitative study of human dynamics (Barabási 2005), through an analysis on the changing process of research interests from the duration perspective, we find that it is not a pure random process, but with more underlying principles.

For scientific researchers, as a key process for learning new knowledge and creating new ideas, searching literatures on the Web can be considered as a process of finding new interests and trying to connect with existing interests that they have in mind (Bransford et al. 2000). Hence, we consider research interests as environmental factors that can be used to refine the search process and help researchers find useful search results that are more relevant to their own background.

The framework of Unifying Search and Reasoning (ReaSearch) proposed in Fensel and van Harmelen (2007) is aimed at removing the scalability barriers of Web-scale reasoning. It emphasizes searching and selecting the most relevant sub-datasets for users before the querying and reasoning process. Based on this framework, concrete selection strategies can be developed. As discussed above, in literature search on the Web, research interests of the users can be considered as contextual constraints that may help to find what the users really want when the original query is vague or there are too many query results that the user has to

wade through to find the most relevant ones. Hence, we propose a concrete method to implement the "ReaSearch" framework proposed in Fensel and van Harmelen (2007), namely, interests-based unification of search and reasoning (I-ReaSearch). In the proposed method, the search and selection of most relevant sub-datasets is based on user interests.

As an application domain of the dynamic and structural characteristics of research interests and "I-ReaSearch", we investigate on how they can be used in literature search on the Web. In order to extract the top N research interests that can be involved in the I-ReaSearch process, we consider evaluating them from the perspectives of cumulative interest value, retained interest value, interest longest duration and interest cumulative duration, as well as group retained interest value. A series of experiments is done based on the DBLP dataset. For the effectiveness of the proposed method, we invite some computer scientists who have several publication in the DBLP system to participate in the evaluation of query results. We also make a comparative study on the scalability of the proposed method. They collectively show that the I-ReaSearch method provides a user centered and practical way to problem solving based on large scale data.

From the standpoint of Active Media Technology (AMT) (Liu 2006), in this paper, the study on the structural and dynamical characteristics of research interests is based on complex networks and human dynamics, which can be considered as an effort towards utilizing information physics to discover and explain the phenomena related to research interests of scientists (Liu 2006). The application of research interests aims at providing scientific researchers best means and best ends in an proactive way for literature search on the Web, based on their own contextual constraints. The study on group retained interests and related applications in query refinement can be considered as utilizing the result of collaboration (in the form of similar research interests) to help researchers find what they really want. This can be considered as an effort towards the application of social interactions in AMT from the Wisdom Web point of view (Liu 2006). Since the study in this paper focuses on the natural characteristics of research interests and how it can be applied to user-centric applications, the overall efforts can also be considered as an endeavor towards "computing and interacting with humans in a human way" (Liu 2006). This paper is extended and refined from three previous papers (Zeng et al. 2009, b, 2010b).

2 Measuring research interests

Measuring research interests may help to get more background information for researchers in order to support their activities on the Web. Nevertheless, not all of them can be measured if the authors do not provide enough information (such as the interests which have not been explicitly shown anywhere). On the other hand, authors' previous publication can be considered as a source where their research interests can be extracted. In this paper, we measure research interests of an author through his/her previous publications. Here we define some parameters to quantitatively measure them.

An interest, more specifically a research interest of an author can be represented as a topic (denoted as terms). Let i, j be positive integers ($i, j \in I^+$), $y_{t(i),j}$ be the number of publications which are related to topic $t(i)$ during the time interval j.

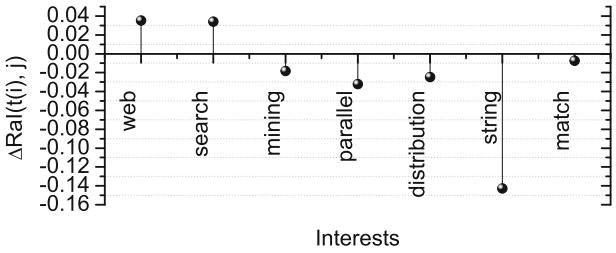


Fig. 1 Relevance ratio of Ricardo' research interests

Cumulative interest, denoted as $CI(t(i), n)$, is used to count the cumulative appear times of $t(i)$ during the considered n time intervals. It can be represented as:

$$CI(t(i), n) = \sum_{j=1}^{n} y_{t(i),j}. \tag{1}$$

It is assumed that the appear times of an interest can be simply added together to reflect a user's overall interest on the specified topic within a time interval.

Ratio of research interest, denoted as $RaI(t(i), j)$, is the ratio between the interest of $t(i)$ and the interest to the set of all m topics that an author is interested in.

$$RaI(t(i), j) = \frac{y_{t(i),j}}{\sum_{i=1}^{m} y_{t(i),j}}. \tag{2}$$

Here we assume that a paper can be categorized into more than one domain which are characterized by terms. Hence, $\sum_{i=1}^{m} y_{t(i),j}$ does not equal to the total number of papers, since one paper may be counted for more than one time. But it equals to the sum of term counts.

Average ratio of research interest, denoted as $avrRaI(m, j)$, is the average value for all the ratio of considered research interests in the time interval j.

$$avrRaI(m, j) = \frac{\sum_{i=1}^{m} RaI(t(i), j)}{m}, \tag{3}$$

where m is the number of considered interests. The relationship between $RaI(t(i), j)$ and $avrRaI(m, j)$ can be denoted as:

$$RaI(t(i), j) = arvRaI(m, j) + \Delta RaI(t(i), j), \tag{4}$$

where $\Delta RaI(t(i), j)$ is the relevance ratio of research interests in the time interval j, which can be calculated as the difference from $RaI(t(i), j)$ to $avrRaI(m, j)$. If $\Delta RaI(t(i), j) < 0$, then the author has a lower interest in $t(i)$ than the average value $avrRaI(m, j)$. If $\Delta RaI(t(i), j) > 0$, then the author has a higher interest in $t(i)$ compared to the average value.

For simplicity, in this paper, we consider single word term to describe research interests. Figure 1 shows the ratio of research interests of the author Ricardo Baeza-Yates based on the DBLP dataset.[1]

[1]The page was visited in Oct. 17th, 2009. A list of filtered words can be found from http://www.wici-lab.org/wici/dblp-sse/Filterwords.txt.

Through a measurement of research interests from publications for scientists, we can have their background concerning research areas automatically. In this section, we focus on an overall statistical analysis through all the years. It seems necessary to study research interests appeared in different time intervals chronologically in order to track its dynamic changing characteristics.

3 Tracking the dynamic shift

Tracking the change of research interests for scientists can identify their recent interests, which can be used to provide more personalized and up-to-date support to their research. In addition, it can help to portray and support understanding on the characteristics of the dynamic process.

To the best of our knowledge, there are few studies related to tracking the dynamic shift of research interests for an author. But there are plenty of studies related to tracking the shift of research interests for a scientific domain. According to current publications, methodologies for identifying the shift of research trends can be divided into three types: the use of contents (e.g. finding frequent words in literature titles) (Erten et al. 2004), the use of citations (Chen 2006; Popescul et al. 2000; Roy et al. 2002), and a combination of the two methods. For the use of citations, one common methodology to identify trends and shift of research interests for a domain is the statistical analysis of word profiles (Chen 2006; Popescul et al. 2000; Roy et al. 2002; Braam et al. 1991; Small and Griffith 1974). The idea is to group a collection of literatures which cite the same set of literatures together and find research front terms by statistical analysis of word profiles (e.g. Top-N most frequent words, Small and Griffith 1974) from those citing literatures (Chen 2006).

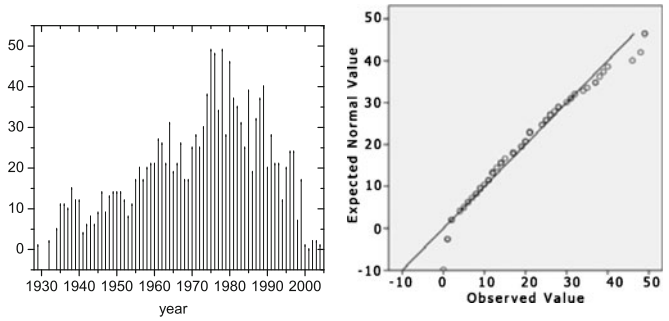
For identification of user interests on the Web, Web page content and click stream analysis has been investigated (Qiu and Cho 2006). In our study, the DBLP dataset only contains author names and publication name related information (no full contents or click stream data provided), hence we concentrate on the word-profile strategy, namely, we use word frequency to detect the dynamic change. In this section, we introduce some parameters to detect the shift of research interests.

Degree of research interest, denoted as $D(t(i), j)$, shows how much is the author interested in the topic $t(i)$ in average during the period of time interval $j = [x_{j-1}, x_j]$ (x_{j-1} and x_j represent the starting time and the ending time of the time interval j):

$$D(t(i), j) = \frac{y_{t(i),j}}{x_j - x_{j-1}}. \tag{5}$$

Based on degree of research interest, one can model the changing process of a research interest in different time intervals. The whole process on the shift of a research interest may be approximate to some kinds of probabilistic distributions.

Figure 2a is an analysis of all the publications of a famous mathematician named "Paul Erdos". Figure 2b shows that all the plots are distributed around a strait line, and by Shapiro wilks measurement, the significance value is 0.058 which is greater than 0.05, hence the distribution of Erdos's publication number over different years follows a normal distribution.



(a) (b)

Fig. 2 An analysis of degree of research interests in different time intervals. **a** Paul Erdos' publication distribution over years based on Erdos' publication collection (1929–1989) and MathSciNet (1990–2004). **b** the Q-Q diagram for **a**

Average degree of research interest, denoted as $avrD(t(i), n)$, is the average value for topic $t(i)$'s degree of research interest in all considered time intervals.

$$avrD(t(i), n) = \frac{\sum_{k=1}^{n} D(t(i), k)}{n}, \tag{6}$$

where $D(t(i), k)$ is the degree of research interest of the topic $t(i)$, $k \in [1, ..., n]$ is a specific time interval. There are n considered time intervals over all.

Relative degree of research interest, denoted as $\delta D(t(i), k)$ is the difference between $D(t(i), k)$ and $avrD(t(i), n)$.

$$\delta D(t(i), k) = D(t(i), k) - avrD(t(i), n). \tag{7}$$

It shows the relationship between $t(i)$'s average degree of the research interest and $D(t(i), k)$ within a specific time interval k.

Degree of research interest growth, denoted as $DG(t(i), j)$, is the growth of research interest degree for $t(i)$ in two consecutive time interval $(j-1)$ and j:

$$DG(t(i), j) = D(t(i), (j-1)) - D(t(i), j). \tag{8}$$

One can compare the research interest growth of different topics through their values of $DG(t(i), j)$. Suppose there are two arbitrary topics $t(i)$ and $t(i')$ from the research interests list of an author. If $DG(t(i), j) > DG(t(i'), j)$, then we say the author's research interest growth in $t(i)$ is higher than $t(i')$ in the time interval j.

Average degree of research interest growth, denoted as $avrDG(t(i), n)$, is the average value of $DG(t(i), j)$ for the topic $t(i)$ in different time intervals.

$$avrDG(t(i), n) = \frac{1}{n} \sum_{j=1}^{n} DG(t(i), j), \tag{9}$$

where n is the total number of considered time intervals.

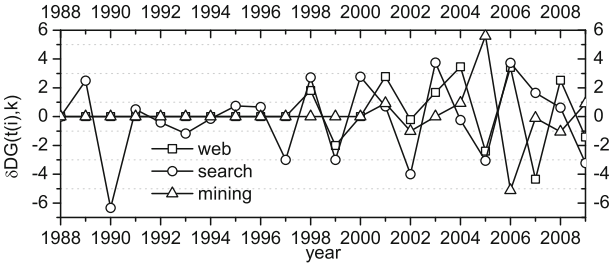


Fig. 3 An analysis of Ricardo's relative degree of research interest growth $\delta DG(t(i), k)$

Relative degree of research interest growth, denoted as $\delta DG(t(i), k)$, is the difference from the research interest growth $DG(t(i), k)$ to the average degree of research interest growth $avrDG(t(i), n)$.

$$\delta DG(t(i), k) = DG(t(i), k) - avrDG(t(i), n). \tag{10}$$

Figure 3 shows Ricardo Baeza-Yates's 3 interests (namely, Web, search, mining) on their relative degree of research interest growth through an analysis of his publication in DBLP. We chose some most interesting topics for him in our study based on a statistical analysis of single-word term frequency from 1987 to 2009.

Weight of a research interest, denoted as $w(t(i), j)$, is the weight of topic $t(i)$ in all the topics appeared in a specified time interval $j = [x_{j-1}, x_j]$.

$$w(t(i), j) = \frac{y_{t(i),j}}{y_j}, \tag{11}$$

where $y_{t(i),j}$ is the number of appeared times of $t(i)$ in the time interval j, and y_j is the total number of events that is related to research interests in the same specified time interval. In our study, an event that is related to research interest is publishing a paper or a book by a specific author.

Suppose j' and j'' are two arbitrary time intervals. For a topic $t(i)$, the corresponding weights of research interest are $w(t(i), j')$ and $w(t(i), j'')$. If $w(t(i), j') > w(t(i), j'')$, then in the time interval j', the research interest in topic $t(i)$ is higher than in j''. Suppose there are two topics $t(i)$ and $t(i')$ in the same time interval j, and their corresponding weights of research interest are $w(t(i), j)$ and $w(t(i'), j)$. If $w(t(i), j) > w(t(i'), j)$, then the author's interest in $t(i)$ is higher than in $t(i')$ in the time interval j.

Figure 4 shows the change of weighted research interests of 3 interests out of 15 that have been selected for investigation. From this figure, we can conclude that in the same period of time, having the same number of publication does not equal to

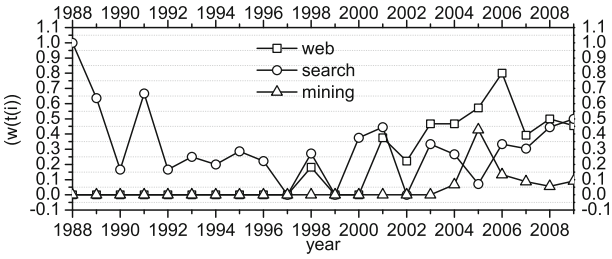


Fig. 4 An analysis of Ricardo's weighted interest change $w(t(i))$

having constant research interest. For example, the author has 2 published papers related to "mining" in the year 2006 and 2009 (excluding edited proceedings), but the interest in this topic has been decreased. That is because the weight of this interest in 2009 is smaller compared to the one in 2006.

4 The acquisition and characteristics of retained interests

As introduced in the above sections, for each author in the scientific community, the research interests are changing all the time, and the above methods and parameters only can help to identify the most recent interests based on the analysis of cumulative interests within a time interval. Nevertheless, the impact of previous interests to the current interests has not been discussed. In this section, we are going to quantitatively investigate on the recent retained interests for authors and their characteristics in the collaborative network environment.

4.1 Obtaining the retained interests

Interests may change over time, and a person may be interested in a topic for a period of time but is likely to loose interest on it as time passes by if it has not appeared in some way for a long time. This phenomenon is very similar to the forgetting mechanism for cognitive memory retention. Hence, we emphasize that the retained interest, which is very related to a user's current interest, can be modeled by using memory retention like functions (Anderson and Schooler 1991). In Zeng et al. (2009b, 2010b), we developed an interest retention model based on a power function that cognitive memory retention follows.

$$RI(t(i), n) = \sum_{j=1}^{n} y_{t(i),j} \times AT_{t(i)}^{-b}, \tag{12}$$

where $T_{t(i)}$ is the duration of the topic $t(i)$ that a user is interested in. In each time interval j, the interest $t(i)$ appears $y_{t(i),j}$ times, and $y_{t(i),j} \times AT_{t(i)}^{-b}$ is the total retention of an interest contributed by the specific time interval.

Here we introduce an example from our previous work (Zeng et al. 2009b, 2010b). We consider a scenario of tracking the authors' research interests which were extracted from their publications. According to the DBLP dataset, the basic time interval considered in our study is a year. Since the retained interests might be related to users' current interests, we use the values from the retained interest model to predict users' current research interests. The values for the parameter "A" and "b" are acquired by maximizing the correlation between the retained interests and the current interests (A is used to minimize the difference between the real interest value and the retained interest value obtained by the proposed function. b is used to control the decaying speed on the lost of interests). According to our previous studies (Zeng et al. 2009b, 2010b), in order to minimize the value of ρ in t-test, the parameters in the retained interest model are chosen as $A = 0.855$ and $b = 1.295$. The value for Spearman's rank order correlation coefficient between the prediction and the real data is $\gamma \approx 0.107$, and for 1-tail t-test, $\rho = 0.237$. The results are, to some extent, close to statistical significance. In order to test the parameters in larger range, in our initial work, we choose all the authors from the SwetoDBLP

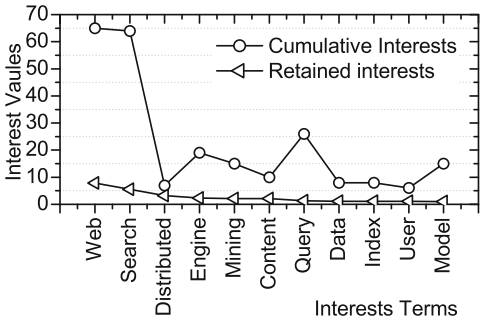


Fig. 5 A comparative study on the cumulative interests and retained interests of the author "Ricardo A. Baeza-Yates" based on the author's publication list from 1987 to 2009

dataset[2] whose number of publications are above 100 (1226 authors in total). Using the retained interest function and relevant parameters introduced above, we extract top 9 interests from the authors' interest lists in the time interval of [2000, 2007] (1226 × 8 sets of data are involved). A comparative study on the actual interests and predicted interests has been done. According to the experimental results, 49.54% of the predictions can match at least 3 interests in the top 9 interests.

Figure 5 provides a comparative study of cumulative interests and retained interests of the author "Ricardo Baeza-Yates". As observed, an interest with relatively high cumulative interest value ($CI(t(i), j)$), does not always have a high retained interest value ($RI(t(i), j)$), such as "query" in the figure. In addition, although some of the interests, such as "distribution" does not have higher cumulative interest values, they may have very high retained interest values since they may be currently, at least most recently interesting to a user.

4.2 Characteristics of retained interests in the collaborative network

In the scientific community, a researcher and his/her collaborators form a collaborative network. His/her research interests may have some relationships with the co-author network since the network contains a group of other researchers who also have some research interests. If they always communicate with each other, in the form of collaboration and coauthoring, etc., their interests may be affected by each others'. Based on the above section, here we give a preliminary study on the relationship of an individual's retained interests and his/her group retained interests.

For a specific interest "t(i)", its group retained interest for a specific author "u", namely "$GRI(t(i), u)$" can be quantitatively defined as:

$$GRI(t(i), u) = \sum_{c=1}^{p} E(t(i), u, c),$$

$$E(t(i), u, c) = \begin{cases} 1 \ (t(i) \in I_c^{topN}) \\ 0 \ (t(i) \in I_c^{topN}) \end{cases} \tag{13}$$

[2]SwetoDBLP dataset is an RDF version of the DBLP dataset. It can be downloaded from http://knoesis.wright.edu/library/ontologies/swetodblp/.

Table 1 A comparative study of top 7 retained interests of a user and his/her group retained interests

User name: Ricardo A. Baeza-Yates

Self retained interests	Value	Group retained interests	Value
Web	7.81	Search (*)	35
Search	5.59	Retrieval	30
Distributed	3.19	Web (*)	28
Engine	2.27	Information	26
Mining	2.14	System	19
Content	2.10	Query (*)	18
Query	1.26	Analysis	14

where $E(t(i), u, c) \in \{0, 1\}$, if the interest $t(i)$ appears both in the top N retained interests of a user and one of his/her coauthors', then $E(t(i), u, c) = 1$, otherwise, $E(t(i), u, c) = 0$. For a specific user "u", there are p coauthors in all, and the group retained interest of "t(i)", namely $GRI(t(i), u)$ is the cumulative value of $E(t(i), u, c)$ based on all his/her coauthors. In a word, group retained interest focuses on the cumulation of ranked interests from a specific user's social network.

We take the author Ricardo A. Baeza-Yates as an example and we examine the relationship between his own retained interests and his group retained interests. Through Table 1 we can find that group retained interests are not the same as, but to some extent related to the user's own retained interests (interesting terms that are marked with "*" are the same). As a step forward, we randomly pick 30 most productive authors from the SwetoDBLP dataset and we calculate their own interests retention and group interests retention. We find that in average, 57% of the top 8 retained interests and group retained interests have overlaps. Namely, if we add all of the coauthors' retained interests together for an author, we observe that many of the top group retained interests are relevant to his/her own retained interests.

In this section, we examined the shift of research interests based on word profiles. It is emphasized that study of emerging trends in a network setting brings more implications, because instead of using first-order word frequency, it provides an understanding of the problem in a graph-theoretical setting (Chen 2006).

5 Building and analyzing the structure of research interests

One can understand the shift and the nature of research interests through analyzing their structures and obtaining corresponding structural characteristics. In this section, we firstly examine the structure of research interests from the network perspective. Then we investigate on the dynamics of these structures from the time perspective.

5.1 Constructing the structure of research interests

From the network perspective, all the research interests can be connected together to form a networked structure. Figure 6 provides some examples of research interests networks based on the author Ricardo A. Baeza-Yates' publication. It shows how research interests (more specifically, cumulative interest $CI(t(i), j)$ in this study) shift as time passes by in the interests network (in this study, we choose the top 8 ranked single-word terms from the year 1991, 1997, 2003, and 2009). The edges show the

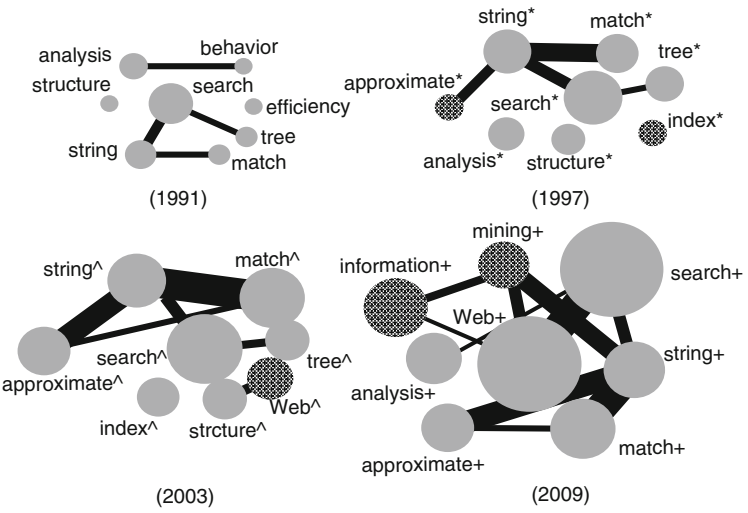


Fig. 6 Ricardos research interest dynamic evolution network from 1991 to 2009. (Based on the author's DBLP publication list, with 232 papers involved). The network is a graph with weighted edges and weighted vertices

connections of two interesting terms (if both of them appear in the same paper title, then a degree of connection is added to them). The width of the edge shows the number of times that two terms are connected together. In the network setting, the research interests are pivotal nodes in the interests network, hence the shift of them shows the major dynamic changing process on the shift of research interests. Some interesting phenomenon has been observed:

(1) Main research interests (pivotal nodes) are dynamically changing all the time. Some pivotal nodes are growing larger in every observed time interval(e.g. search), which may be due to growing interest in the specific topics, and some of them disappeared from the first 8 pivotal nodes (e.g. tree, behavior, index), which may be due to the lost of interest on corresponding topics. Meanwhile, some new research interests emerged (the ones that are marked with decorative patterns, e.g. "Web", "mining").
(2) Most of the top research interests are connected with each other. It indicates that for an author, his/her research interests are not isolated, instead, they are highly relevant and co-occur with each other frequently.
(3) The figure shows that relations among research interests vary as time passes by (e.g. the connections between "Web" and "search"). If the author has interest in working on the unification of two related topics, then connections between them will grow stronger (the degree of connections is shown as the thickness of the edges in the interests networks).
(4) Many top research interests (pivotal nodes) remain active in the interest network (e.g. search, analysis, match), this can be explained from the perspective of preferential attachment in network science theory, namely, new nodes in a network prefer and are likely to be connected with the pivotal nodes (Barabási 2002).

5.2 Analyzing the structure of research interests

The structure which is composed of all research interests of an author is with some characteristics. In this paper, we will study two types of characteristics, namely, degree characteristics and timing characteristics of research interests.

5.2.1 Degree characteristics of research interests

In order to examine the degree characteristics of an author's research interests, we consider to choose all the time intervals that the author has publications and add the weight of research interests ($w(t(i), j)$) together. In this case, the cumulative value of $w(t(i), j)$ is the degree of the node $t(i)$ in the interests network (i.e. the cumulative interest value in the whole time intervals).

Here we examined three authors' degree characteristics of research interests. Namely, Leonhard Euler, Paul Erdos, and Ricardo Baeza-Yates respectively. Figure 7 indicates that concerning the degree of the node $t(i)$ for each single-word term, only a few of them are high, and most of them are relatively low. It shows that research Interests for these authors are with power-law distribution. The slope in corresponding figures are: -1.62 ± 0.15 (Euler), -1.15 ± 0.07 (Erdos), and -1.33 ± 0.14 (Ricardo). Which shows that in scientific research, we may approximately consider that the slope value may be very close for different authors. Although the slope value is not that close as people observed in other human activities, such as mail correspondence (with slope value 1.5, Oliveira and Barabási 2005).

The degree characteristics of research interests show that there are some major research interests (as pivotal nodes), which play central roles and are of vital importance in the research interests network structure (Barabási 2002). These major research interests may have effect on other interests in the network.

5.2.2 Timing characteristics of research interests

Traditionally human activities are approximately modeled using poisson process, which is based on a hypothesis of their random distribution in time (Barabási 2005; Greene 1997; Haight 1967; Reynolds 2003). Recent findings emphasize that consider from the time perspective, many human activities (e.g. email and short message

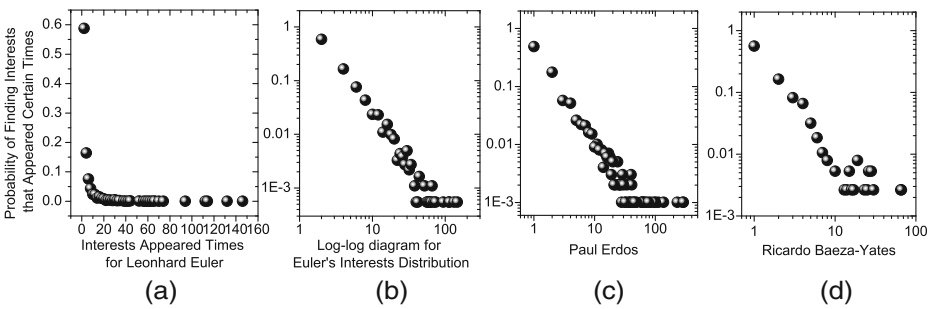


Fig. 7 Power-law distribution on weights of research interests for Leonhard Euler (Publication list is from Euler's Archive), Paul Erdos (publication list is from Erdos' publication collection and MathSciNet), and Ricardo Baeza-Yates (publication list is from DBLP)

sending, online clicking of web pages, making calls, financial commerce, etc.) follow power-law distribution (Barabási 2005; Dezso et al. 2006; Han et al. 2008; Masoliver et al. 2003). The results indicate that for human activities, instead of pure random processes, there might be more underlying principles. Scientific research is a typical human activity, and the process on the shift of research interest is closely related to time. To the best of our knowledge, there is few study on timing statistical characteristics on the shift of research interests.

A specific research interest may appear in different time intervals, its distribution characteristics may not be the same. For those, which keep a relatively steady interest, may have a poisson distribution. For those, which have a gradual increase and then have a gradual decrease, may follow gaussian distribution. For those, which have a burst of research interest and then reduce sharply to a low interest and last for a relatively long time, some time later may be back to another burst, may follow power-law distribution. Nevertheless, when we put all the research interests together and investigate on the statistical characteristics, some interesting distribution can be observed.

The process on the shift of research interest is to some extent different from email sending, online clicking of web pages, etc., which have actions one by one. An author is likely to have more than one research interests during a time interval and each of them doesn't come one after another, instead, they may exist at the same time. Authors publish results in different time intervals. It motivates us to investigate on the statistical characteristics of interests duration.

Interest Duration, denoted as $ID(t(i))$, is used to represent the duration of the interest $t(i)$ between it appears and disappears. If the interest $t(i)$ appears several times in one basic time interval(e.g. a month, a year, etc.), it will be counted just once. At least two parameters can be used to investigate the characteristics of interest duration, namely, interest longest duration and interest cumulative duration.

Interests Longest Duration, denoted as $ILD(t(i))$, is used to represent the longest duration of the interest $t(i)$:

$$ILD(t(i)) = \max\left(ID(t(i))_q\right), \tag{14}$$

where $q \in I^+$, $ID(t(i))_q$ is the interest duration when $t(i)$ discretely appears (two different interest durations should be separated, not continuous.) for the qth time.

Interests Cumulative Duration, denoted as $ICD(t(i))$, is used to represent the cumulative duration of the interest $t(i)$. It shows how long the interest has appeared:

$$ICD(t(i)) = \sum_{q=1}^{q'} \left(ID(t(i))_q\right), \tag{15}$$

where $q \in I^+$ is used to represent the qth discrete appearance of the interest $t(i)$, and q' is the total discrete appearance times of the interest $t(i)$.

Figure 8a is an analysis of Ricardo Baeza-Yates's $ILD(t(i))$. Notice that there are some large spikes in the plot, corresponding to very long $ILD(t(i))$ for some research interests. It indicates that the interest longest duration distribution of research interests is a non-poisson process. Figure 8b is an analysis on the probability of having n research interests whose longest interest duration is a fixed time interval (τ). This statistical distribution is best approximated as:

$$P(\tau) \approx \tau^{-\alpha}, \tag{16}$$

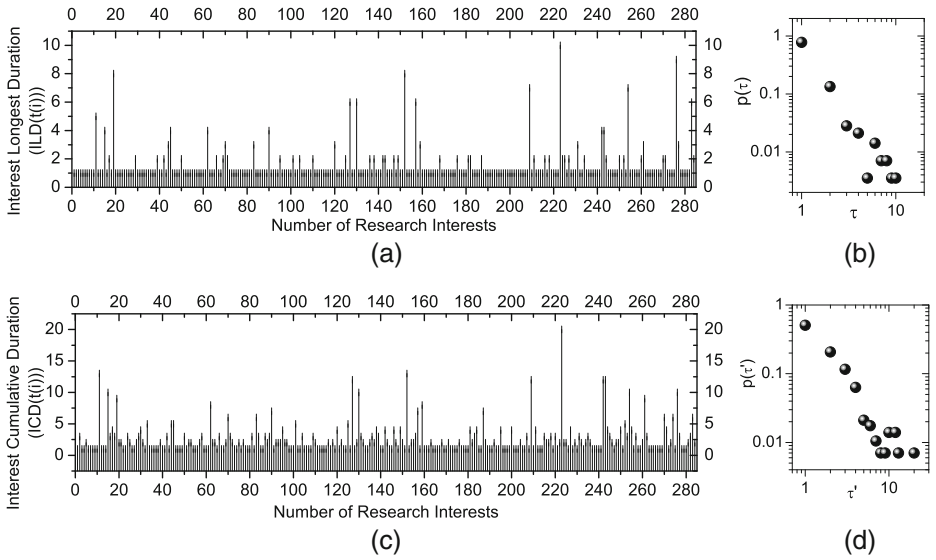


Fig. 8 Ricardo's research interest longest duration and cumulative duration statistics

where $\alpha \simeq 1.64$ (the slope of the solid line in the log-log plot is -1.64), which indicates that an author's research interest shifting pattern has a power-law character: for most research interests, they will not last for a long time, and for a relatively small number of research interests, they may last comparatively much longer.

Figure 8c is an analysis of Ricardo's $ICD(t(i))$. We can observe similar phenomenon as in Fig. 8a, that there are some large spikes in the plot, corresponding to very long $ICD(t(i))$ for some research interests. As shown in Fig. 8d, the statistical distribution on the value of $ICD(t(i))$ can be best approximated as:

$$P(\tau') \approx \tau'^{-\alpha'}, \tag{17}$$

where $\alpha' \simeq 2.30$ (the slope of the solid line in the log-log plot is -2.30), τ is the number of interests whose $ICD(t(i))$ are equal to each other. The figure indicates that the $ICD(t(i))$ distribution also follows the power-law. Most research interests have a small number of years of appearance, while some of the research interests appear in many observed years.

Figure 8a and c shows that the $ILD(t(i))$ and $ICD(t(i))$ for the same $t(i)$ do not always consistent (the x-axises of these two figures share the same corresponding interests), namely, one research interest may have appeared in many years, hence it has relatively longer $ICD(t(i))$, but has a relatively shorter $ILD(t(i))$, which shows that an author may not have a continuous interest in a topic but has interest working on it after some break (if he/she finds some unsolved interesting problems).

The reason why the distribution of $ILD(t(i))$ follows power-law distribution can be explained as follows: (1) Compared to those more specific ones, most of the interests which last for a relatively long time are more general. They seem to have more unsolved problems. (2) The interests which last for a relatively long time are related to many specific interests, namely, they are correlated events.

The reason why the distribution of $ICD(t(i))$ follows a power-law can be explained as follows: (1) As shown in Fig. 8, although the rank order of $ILD(t(i))$ is not consistent with the $ICD(t(i))$ all the time, it is very related. And if a research interest has a relatively long $ILD(t(i))$, its' probability of having a relatively long $ICD(t(i))$ is very high. (2) If an author always find some unsolved interesting problems after a break, he/she is likely to come back to the topic, and in this case, this research interest may have relatively longer $ICD(t(i))$. (3) According to the statistical results, In most cases, if an author left a topic, it is probably not going to come back. These research interests have relatively smaller number of appearance times.

We have analyzed all the authors' interests values based on the DBLP dataset using the introduced models (namely, the cumulative interests, retained interests, interests longest duration, interests cumulative duration), and we have described authors' interests values in an RDF file.[3]

6 A framework of unifying search and reasoning with user interests

The "ReaSearch" approach proposed in Fensel and van Harmelen (2007) is aimed at solving the problem of scalability for Web-scale reasoning. It's core philosophy is to select appropriate subsets of semantic data for reasoning, and it tries to solve the scalability issue by incomplete reasoning since the dataset that is acquired from the Web itself is incomplete anyway.

From the network theory perspective, the process of investigating unexplored topics can be considered as adding new nodes to the interests network. By the phenomenon of preferential attachment which has been briefly discussed in Section 5.1, we can predict that unexplored topics are likely to be connected with major research interests (namely, the pivotal nodes) in the interests network. In addition, bridging a new topic with familiar ones can help to understand the new and is convenient for human to learn (Bransford et al. 2000). Hence, research interests can be considered as contextual information for literature search on the Web. In this paper, we propose to unify search and reasoning based on user interests. Following the notion in Fensel and van Harmelen (2007), we name the efforts as "I-ReaSearch", which means unifying reasoning and search with Interests. The process of I-ReaSearch can be described as the following rule:

$$hasInterests(U, I), hasQuery(U, Q), executesOver(Q, D), \neg contains(Q, I) \rightarrow IReaSearch(I, Q, D),$$

where $hasInterests(U, I)$ denotes that the user "U" has a list of interests "I" and can be acquired, $hasQuery(U, Q)$ represents that there is a query input "Q" by the user "U", $executesOver(Q, D)$ denotes that the query "Q" is executed over the dataset "D", $\neg contains(Q, I)$ represents that the query "Q" does not contain the list of interests "I", $IReaSearch(I, Q, D)$ denotes that by utilizing the interests list "I" and the query "Q", the process of unifying search and reasoning is taken on the dataset "D".

[3]The RDF version of the DBLP authors' interests dataset has been released through http://wiki.larkc.eu/csri-rdf.

This approach implements the idea of "refining querying by using rules" proposed in Berners-Lee and Fischetti (1999). Two methods are proposed for "I-ReaSearch", namely, the implementations of $IReaSearch(I, Q, D)$ are with two directions. The first one is user interests based query refinement, and the second is interleaving selection and reasoning (now focusing on querying) based on user interests. Both of these strategies utilize user interests as the contextual information, but the processing mechanisms are different.

6.1 Interests-based query refinement

For the strategy of user interests based Query Refinement, it adds more constraints to the query according to the user interests extracted from some historical sources (such as previous publication, visiting logs, etc.). The process can be described by the following rule:

$$hasInterests(U, I), hasQuery(U, Q), executesOver(Q, D), \neg contains(Q, I) \rightarrow refinedAs(Q, Q'), contains(Q', I), executesOver(Q', D).$$

In this rule, $refinedAs(Q, Q')$ represents that the original query "Q" is refined by using the list of interests as "Q'". $contains(Q', I)$ denotes that "Q'" contains the list of interests "I". $executesOver(Q', D)$ represents that the refined query "Q'" executes over the dataset "D". Namely, "$refinedAs(Q, Q'), contains(Q', I), executesOver(Q', D)$" implements $IReaSearch(I, Q, D)$ in the I-ReaSearch general framework.

Based on the upper rule, we emphasize that this approach utilizes the user context to provide a rewritten query so that more relevant results to the specific user's interests can be acquired.

6.2 Querying with interests-based selection

For the strategy of querying with interests-based selection, it emphasize a selection step before querying and reasoning on the data, since user interests might help to find a more relevant sub dataset for each specific user compared to querying on the whole. The process can be described by the following rule:

$$hasInterests(U, I), hasQuery(U, Q), executesOver(Q, D), \neg contains(Q, I) \rightarrow Select(D', D, I), executesOver(Q, D'),$$

where "Select(D', D, I)" represents the selection of a sub dataset "D'" from the original dataset "D" based on the interests list "I", and $executesOver(Q, D')$ represents that the query is executed over the selected sub dataset "D'". Namely, in the upper rule, "$Select(D', D, I), executesOver(Q, D')$" implements $IReaSearch(I, Q, D)$ in the I-ReaSearch general framework.

7 Interests-based reasearch from different perspectives

When the query is vague/incomplete, research interests can serve as constraints that can be used to refine these queries. They can be evaluated from various perspectives and each perspective reflects one unique characteristics of research

interests. When the user is not satisfied with one perspective, the interests list that is used for interests-based ReaSearch can be changed. The process of interests-based ReaSearch can be described by the following rule:

$$IReaSearch(I, Q, D), \neg satisfies(U, R) \rightarrow IReaSearch(I', Q, D),$$

where $IReaSearch(I, Q, D)$ denotes that the ReaSearch process is based on the interest list "I" and the query "Q" on the dataset "D". $\neg satisfies(U, R)$ denotes that the user does not satisfy with the query results from $IReaSearch(I, Q, D)$. $IReaSearch(I', Q, D)$ denotes that the interest list is changed from "I" to "I'" and the ReaSearch process is based on the new interest list"I'".

Based on the above discussion of this study, we can have various perspectives to acquire a list of top N research interests. In the following section, we will focus on 4 perspectives, namely, the cumulative interest introduced in Section 2, the retained interest introduced in Section 4, the interest longest duration, and the interest cumulative duration introduced in Section 5.2.

As an illustrative example, Table 2 provides a comparative study of an author's top 7 interests with the biggest retained interest values, with the biggest interest longest duration and the interest cumulative duration values. As shown by the table, the ranking of the interests are different when we investigate them from different perspectives. Hence, when we consider using research interests to refine literature search, various results can be obtained by using top N interests from these perspectives. Table 3 shows a partial comparative study of search results using a vague query "intelligence" and implicit constraints from various interest lists are added to the original query. Based on these three perspectives, different search results are selected out and provided to users to meet their diverse needs (In this partial list of results, literatures which contain the query keywords and constraints from research interests are selected out and ranked to the top. As an illustrative example, in each list, our system shows the first search results that are obtained according to the constraints from each of the research interests).

The results in List 1–3 of Table 3 are acquired based on the context of the user's own interest lists. All of these refined query results are based on the assumption that a user is willing to share his/her real identity so that the interest lists can be acquired. If the user does not want to log in the query system using the real identity, as discussed in Section 4.2, we still can get part of his/her interests on the basis that the user is willing to tell who have been collaborated with him/her in the form of

Table 2 Top 7 interests with the biggest retained interest values (RI), interest longest duration (ILD) values or interest cumulative duration (ICD) values

RI		ILD		ICD	
Web	7.81	Search	10	Search	20
Search	5.59	Web	9	Retrieval	14
Distributed	3.19	Text	8	Algorithm	13
Engine	2.27	Match	8	Text	13
Mining	2.14	Approximate	8	Match	13
Content	2.10	Retrieval	7	Query	12
Query	1.26	Query	7	String	12

User name: Ricardo A. Baeza-Yates

Table 3 Search refinement using the top 7 interests that have the biggest retained interest values, interest longest duration, or interest cumulative duration

Name	Ricardo A. Baeza-Yates
Query	Intelligence
List 1	With the top 7 interests that have the biggest retained interest values: Web, Search, Distributed, Engine, Mining, Content, Query * SWAMI: Searching the **Web** Using Agents with Mobility and **Intelligence**. * Moving Target **Search** with **Intelligence**. * Teaching **Distributed** Artificial **Intelligence** with RoboRally. * Prototyping a Simple Layered Artificial **Intelligence Engine** for Computer Games. * Web Data **Mining** for Predictive **Intelligence**. * **Content** Analysis for Proactive **Intelligence**: Marshaling Frame Evidence. * Efficient XML-to-SQL **Query** Translation: Where to Add the **Intelligence**?
List 2	With the top 7 interests that have the biggest interest longest duration: Search, Web, Text, Match, Approximate, Retrieval, Query * Moving Target **Search** with **Intelligence**. * SWAMI: Searching the **Web** Using Agents with Mobility and **Intelligence**. * **Text**-Based Systems and Information Management: Artificial **Intelligence** Confronts Matters of Scale. * A Multilayer Perceptron Solution to the **Match** Phase Problem in Rule-Based Artificial **Intelligence** Systems. * A New Swarm **Intelligence** Coordination Model Inspired by Collective Prey **Retrieval** and Its Application to Image Alignment. * Efficient XML-to-SQL **Query** Translation: Where to Add the **Intelligence**?
List 3	With the top 7 interests that have the biggest interest cumulative duration: Search, Retrieval, Algorithm, Text, Match, Query, String * Moving Target **Search** with **Intelligence**. * A New Swarm **Intelligence** Coordination Model Inspired by Collective Prey **Retrieval** and Its Application to Image Alignment. * Artificial **intelligence** diagnosis **algorithm** for expanding a precision expert forecasting system. * **Text**-Based Systems and Information Management: Artificial **Intelligence** Confronts Matters of Scale. * A Multilayer Perceptron Solution to the **Match** Phase Problem in Rule-Based Artificial **Intelligence** Systems. * Efficient XML-to-SQL **Query** Translation: Where to Add the **Intelligence**?
List 4	With group retained interests constraints (top 7 results): Search, Retrieval, Web, Information, System, Query, Analysis * Moving Target **Search** with **Intelligence**. * A New Swarm **Intelligence** Coordination Model Inspired by Collective Prey **Retrieval** and Its Application to Image Alignment. * SWAMI: Searching the **Web** Using Agents with Mobility and **Intelligence**. * Building an **information** on demand enterprise that integrates both operational and strategic business **intelligence**. * An Explainable Artificial **Intelligence System** for Small-unit Tactical Behavior. * Efficient XML-to-SQL **Query** Translation: Where to Add the **Intelligence**?

co-authoring papers. In List 4 of Table 3, the user Ricardo A. Baeza-Yates' top 7 group retained interests are used to refine the original query that the user provides. One can see that how the group interests serve as environmental factors that affect the search refinement process and help to get more relevant search results.

Through the above studies, one can get a preliminary idea on how the research interests evaluated from various perspectives can be involved in the ReaSearch process and help the researchers get more relevant query results for further investigations.

8 Evaluation of interests-based ReaSearch

In Section 7, we only provide some examples on how the research interests can be used as environmental factors to affect the query process, but we have not evaluated the effectiveness of the Interests-Based ReaSearch method. In this section, we evaluate the proposed method from the perspective of user satisfaction and scalability.

In order to show the effectiveness of the interests-based ReaSearch method, we invited 10 computer scientists who have at least 5 papers listed in the DBLP dataset to participate in the evaluation. They are required to search for "intelligence" in the DBLP Search Support Engine[4] that we developed based on the SwetoDBLP dataset (Aleman-Meza et al. 2007). Three lists of query results are provided to each of them. The first list is acquired based on the original query input by the user, the second list is with query refinement based on the user's top 7 retained interests, and the third list is with query refinement based on the user's top 7 group retained interests. They are required to judge which list of results they prefer. Through an analysis of the feedbacks from the users, we find that: 100% of these authors feel that the refined search results using user's own retained interests and group retained interests are much better than the result list which does not have any refinement. 100% of them feel that the satisfaction degree of the two refined result lists are very close. 90% of them feel that refined results by the users' own retained interests are better than others. 10% of them feel refined results by group retained interests are better than others.

The refined list with the authors' own retained interests is supposed to be the best one, since the query constraints are all most related topics that the users are interested in. Since the group retained interests of a specific user have some overlap with the user's own retained interests, the refined query results with group retained interests are also welcome and considered much better than the one without any refinement. It indicates that if one's own interests can not be acquired, his/her collaborators' interests also could implicitly help to refine the search process and results.

From the perspective of Scalability, we have made a comparative study on the query effectiveness among three different strategies:

- Strategy 1: Query based on the original user inputs. This strategy only uses the query input by the user and does not contain any refinement.
- Strategy 2: Interests-based query refinement. The query refinement strategy provides a rewritten SPARQL query based on acquired interests of a user (as introduced in Section 6.1).

[4] http://www.wici-lab.org/wici/dblp-sse

– Strategy 3: Querying with Interests-based selection. This strategy selects relevant sub dataset based on user interests acquired from contexts, then the original user input query is performed on the selected sub dataset (as introduced in Section 6.2).

We take the SwetoDBLP dataset which is divided into 22 sub datasets. We evaluate the 3 implemented strategies by using these datasets at different scales. A comparative study is provided in Fig. 9. Two users are taken as examples, namely Frank van Harmelen and Ricardo Baeza-Yates. Top 9 retained interests for each of them are acquired based on the retained interest function (formula 12) and used to unify the search and reasoning process. The above three different kinds of querying strategies are performed on the gradually growing dataset (each time 2 subsets with almost the same size are added, around 55M for each, and 1.08G in total).

As shown in Fig. 9, since "interests-based query refinement (strategy 2)" takes more constraints compared to strategy 1, it requires more processing time, and as the size of the dataset grows, the processing time grows very rapidly, which means that this method does not scale well if we just consider the cost of time. Although this method takes more time, the quality of the query results is much better than strategy 1.

Since "querying with interests-based selection (strategy 3)" selects relevant sub dataset in advance, the required query time is significantly reduced, and as the size of the dataset grows, compared to strategy 1 and strategy 2, the query time is always comparatively shorter and does not increase equally fast. At the same time, the quality of the query results is the same with strategy 2. Hence, this method scales better.

As a step forward, for strategy 2 and strategy 3, we examined the impact on the number of constraints in a query so that one can immediately see the necessity of having a balance among query refinement and processing time. As a follow up, Fig. 10 shows the query processing time with 3,6, and 9 interests constraints from the user "Frank van Harmelen". After analyzing this figure, we can conclude that the number of constraints in the query is positively correlated with the query processing

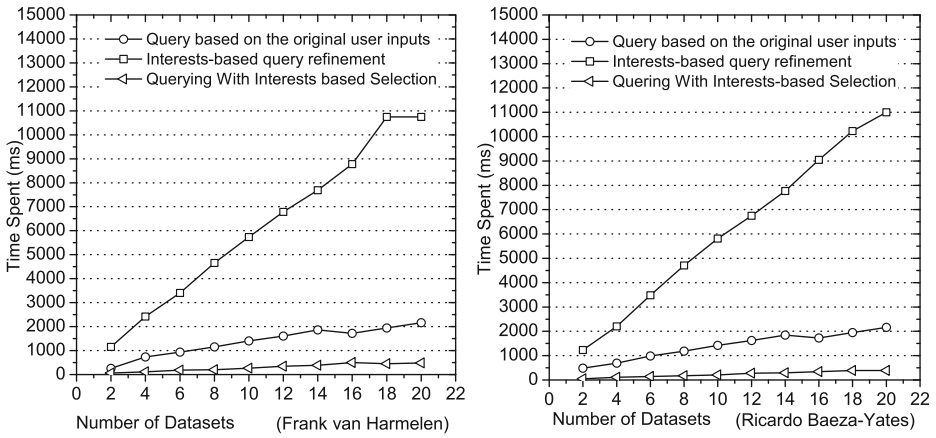


Fig. 9 Scalability on query time for three different strategies

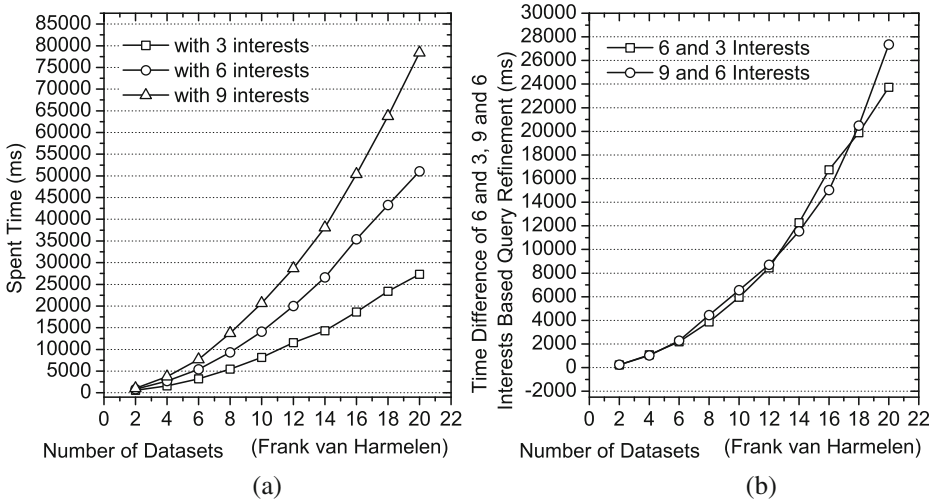


Fig. 10 Interests based refined query time with 3, 6, 9 interests (**a**) and a comparative study of their time difference (**b**)

time. Hence, even if strategy 2 and 3 yield better results compared to strategy 1, one should be cautious about adding too many constraints to the original query each time, since the query processing time might increase rapidly.

Figure 11 gives an analysis on the ratio of spent time between the strategy 1 and strategy 3. As shown in the figure, with the dataset growing to larger scales, the ratio of spent time is not changing too much (Take Frank van Harmelen as an example, the values of the ratio are always between 0.064 and 0.101, while for Ricardo Baeza-Yates, the ratio is between 0.097 and 0.201), and the change on the ratio of query results also stays within a small range (For Frank van Harmelen, the ratio of query results changes between 0.256 and 0.287, and for Ricardo Baeza-Yates, the ratio

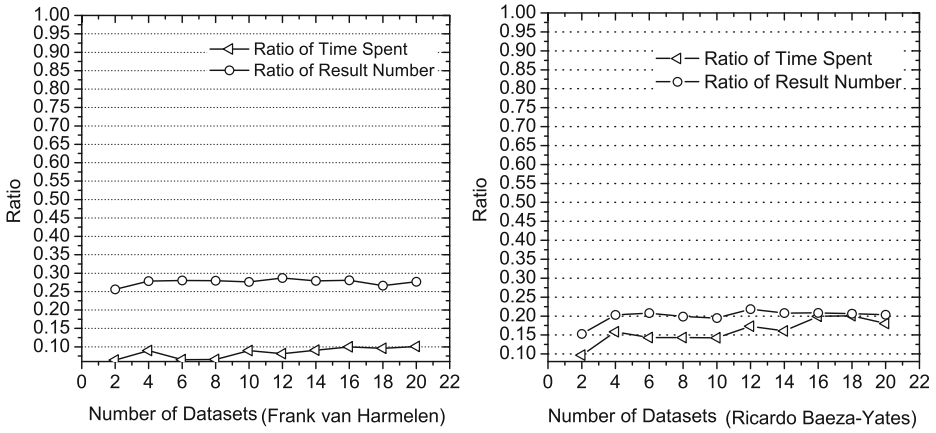


Fig. 11 A comparative study of spent time and recall

changes between 0.153 and 0.219). Namely, the ratio of spent time and the ratio of query results from the 2 strategies are relatively stable, which indicates that strategy 3 scales well. (If the ratio of spent time or the ratio of query results changes a lot as the dataset grows, then the method does not scale well).

At the same time, we should emphasize that although the ratio of results is not high, the query results from strategy 2 and strategy 3 are much closer to the user needs and backgrounds. As indicated by the user feedbacks in this section, the quality of the query results from strategy 2 and 3 are much better.

9 Conclusion and future work

This paper concentrates on the study of interests, more specifically research interests, and their applications on Web literature search. We have provided an understanding for the dynamic and structural characteristics of research interests. For the dynamic perspective, in this paper, we have discussed some preliminary methods for tracking the dynamic changing process of research interests. Within the changing process, the retained interest provides some relevant information on researchers' current interests, hence, we proposed a retained interests model based on the forgetting mechanism in Cognitive Science so that one can quantitatively evaluate how many historical interests have left in users' current interests. For the structural perspective, from the viewpoint of network theory, we have investigated the statistical distribution on the structure and evolution process of interests network and provided some basic understanding on the evolution characteristics of the interests network.

This study not only intends to provide a preliminary understanding on the nature and models of research interests, but also aims at applying related results as environmental, contextual basis to provide better services for researchers during the process of literature search on the Web. In this paper, based on the framework of unifying search and reasoning (ReaSearch), we proposed the method of Interests-based ReaSearch (I-ReaSearch), which aims at unifying search and reasoning based on user interests for Web problem solving. Then under the context of scientific literature search, we provided some illustrative examples on how to unify search and reasoning using acquired interests from different perspectives (namely, the perspectives of retained interest, interest longest duration and interest cumulative duration) considering their timing and structural characteristics. Two concrete strategies have been developed for I-ReaSearch, namely, interests-based query refinement and querying with Interests-based selection. By adding user interests as the context for the query, interests-based query refinement yields better query results compared to an unrefined query, but at the same time takes more processing time. It does not scale well since the processing time increases too quickly. These scaling problems, however, can be taken care of by applying querying with interest-based selection. The results from these two strategies are equivalent, but the latter requires much less query processing time.

In order to enlarge the statistical significance and study each interest in a more general way, we only consider research interests that are single word terms (For example, in our study, we extract 'Web" from "Web search", "Web mining", "Semantic Web", etc. to show that the scientist is generally interested in "Web". In this way, more terms can be acquired to improve the significance when we investigate

on the statistical distribution of interests). After finding these characteristics, for more effective search refinement, we are going to consider multiple word terms. For scientists, their research interests are not only related to themselves, but also have close relationship with their collaborators (e.g. research partners and coauthors) and related academic communities. In future studies, we are going to investigate on how the collaborators and research communities affect the changing process of researchers' interests. For example, we are going to study on how emerging trends, triggering events in a field affect scientists' future research. From the evaluation perspective, in this paper, we only illustrated how the obtained interests can be used for search refinement based on vague queries, and have not done enough results evaluation by real scientists. Hence, we are going to invite more scientists to evaluate which kinds of refinement from research interests they would like and whether these refinement strategies can help during their literature search process on the Web.

From the Active Media Technology (AMT) point of view, following the outline of AMT from the standpoint of Wisdom Web (Liu 2006), in this paper, we try to apply information physics related theories (including complex networks and human dynamics) to investigate on research interests and utilize the analysis results from the observed phenomena to develop user-centered Web applications. The proposed methods and related results show that our aim is to discover best, or at least efficient means (Liu 2006) to help users find what they really want. From the Wisdom Web point of view, this can be considered as some efforts on knowledge retrieval through providing user-centric supporting functionalities (Yao 2002; Yao and Yao 2003; Yao et al. 2007; Hoeber 2008). In this way, one may get practical wisdom of actively serving scientists' working on research (Liu 2006).

Acknowledgements This study is supported by the European Commission under the 7th framework programme, Large Knowledge Collider (FP7-215535). The author would like to thank Yiyu Yao for his constructive discussion on user interests based knowledge retrieval, Rui Guo and Chao Gao on their useful comments on network theory which are used for interpreting the phenomenon observed in this study, Jian Yang for his suggestions on measurement of research interests.

References

Aleman-Meza, B., Hakimpour, F., Arpinar, I. B., & Sheth, A. P. (2007). Swetodblp ontology of computer science publications. *Journal of Web Semantics, 5*(3):151–155.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396–408.

Barabási, A. L. (2002). *Linked: How everything is connected to everything else and what it means for science, business and everyday life* (1 ed.). Perseus Publishing.

Barabási, A. L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature, 435*, 207–211.

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the web: The original design and ultimate destiny of the world wide web by its inventor*. HarperSanFrancisco.

Braam, R. R., Moed, H. F., & Raan, A. F. J. v. (1991). Mapping of science by combined co-citation and word analysis: II. Dynamical aspects. *Journal of the American Society for Information Science, 42*(4), 252–266.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. National Academy Press.

Chen, C. M. (2006). Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Dezso, Z., Almaas, E., Lukács, A., Rácz, B., Szakadát, I., & Barabási, A. L. (2006). Dynamics of information access on the web. *Physical Review E, 73*(066132), 1–6.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology hermann ebbinghaus*. Teachers College, Columbia University.
Erten, C., Harding, P. J., Kobourov, S. G., Wampler, K., & Yee, G. (2004). Exploring the computing literature using temporal graph visualization. In *Proceedings of the 2004 SPIE Conference on visualization and data analysis* (Vol. 5295, pp. 45–56).
Fensel, D., & van Harmelen, F. (2007). Unifying reasoning and search to web scale. *IEEE Internet Computing, 11*(2), 96, 94–95.
Greene, J. H. (1997). *Production and inventory control handbook* (3 ed.). New York: McGraw-Hill.
Haight, F. A. (1967). *Handbook of the Poisson distribution*. New York: Wiley.
Han, X. P., Zhou, T., & Wang, B. H. (2008). Modeling human dynamics with adaptive interest. *New Journal of Physics, 10*(073010), 1–8.
Hoeber, O. (2008). Web information retrieval support systems: The future of web search. In *Proceedings of the 2008 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (Vol. 3, pp. 29–32).
Liu, J. M. (2006). Active media technologies (amt) from the standpoint of the wisdom web. In Y. Li, M. Looi, & N. Zhong (Eds.), *Proceedings of the 4th international conference on active media technology (AMT 2006). Frontiers in artificial intelligence and applications* (Vol. 138, pp. 3–6). IOS Press.
Masoliver, J., Montero, M., & Weiss, G. H. (2003). Continuous-time random-walk model for financial distributions. *Physical Review E, 67*(021112), 1–10.
Oliveira, J. G., & Barabási, A. L. (2005). Darwin and Einstein correspondence patterns. *Nature*, *437*, 1251.
Popescul, A., Flake, G. W., Lawrence, S., Ungar, L. H., & Giles, C. L. (2000). Clustering and identifying temporal trends in document databases. In *Proceedings of the 2000 IEEE advances in digital libraries* (pp. 173–182).
Qiu, F., & Cho, J. (2006). Automatic identification of user interest for personalized search. In *Proceedings of the 2006 international world wide web conference*.
Reynolds, P. (2003). *Call center staffing*. Lebanon, Tennessee: The Call Center School Press.
Roy, S., Gevry, D., & Pottenger, W. M. (2002). Methodologies for trend detection in textual data mining. In *Proceedings of the 2002 text mining workshop at the second SIAM conference on data mining*.
Shneiderman, B. (2008). Science 2.0. *Science, 319*, 1349–1350.
Small, H. G., & Griffith, B. C. (1974). The structure of scientific literatures: I. Identifying and graphing specialties. *Science Studies, 4*, 17–40.
Yao, Y. Y. (2002). Information retrieval support systems. In *Proceedings of the 2002 IEEE international conference on fuzzy systems* (pp. 773–778).
Yao, J. T., & Yao, Y. Y. (2003). Web-based support systems (reprint from wss'03). In *Proceedings of the 2004 international workshop on web-based support systems* (pp. 1–5).
Yao, Y. Y., Zeng, Y., Zhong, N., & Huang, X.J. (2007). Knowledge retrieval (kr). In *Proceedings of the 2007 IEEE/WIC/ACM international conference on web intelligence* (pp. 729–735).
Zeng, Y., Ren, X., Qin, Y. L., Zhong, N., Huang, Z. S., & Wang, Y. (2009). Social relation based scalable semantic search refinement. In *The 1st asian workshop on scalable semantic data processing (AS2DP)*.
Zeng, Y., Yao, Y. Y., Zhong, N. (2009b). Dblp-sse: A dblp search support engine. In *Proceedings of the 2009 IEEE/WIC/ACM international conference on web intelligence* (pp. 626–630).
Zeng, Y., Wang, Y., Huang, Z. S., Damljanovic, D., Zhong, N., et al. (2010a) User interests: Definition, vocabulary, and utilization in unifying search and reasoning. In *Proceedings of the 2010 international conference on active media technology* (pp. 98–107). Springer.
Zeng, Y., Zhong, N., Wang, Y., Qin, Y. L., Huang, Z. S., Zhou, H. Y., et al. (2010b). User-centric query refinement and processing using granularity based strategies. *Knowledge and Information Systems*. doi:10.1007/s10115-010-0298-8.
Zeng, Y., Zhou, E. Z., Qin, Y. L., & Zhong, N. (2010b). Research interests: Their dynamics, structures and applications in web search refinement. In *Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence* (pp. 639–646). IEEE Computer Society.