Research Interests : Their Dynamics, Structures and Applications in Web Search Refinement

Yi Zeng[1], Erzhong Zhou[1], Yulin Qin[1,2], Ning Zhong[1,3]
[1]International WIC Institute, Beijing University of Technology, Beijing, P.R. China
[2] Department of Psychology, Carnegie Mellon University, Pittsburgh, U.S.A
[3] Department of Life Science and Informatics, Maebashi Institute of Technology, Maebashi, Japan
yzeng@emails.bjut.edu.cn, zezbj@yahoo.cn, yq01@andrew.cmu.edu, zhong@maebashi-it.ac.jp

***Abstract*—For most scientists, their research interests are dynamically changing all the time. Through an analysis of research interests, we find that all the changes are with some characteristics. Plus, the research interests in the dynamic changing process are not isolated, instead, they are interconnected as a whole to form a holistic structure. We introduce some measurement parameters to track and detect the evolution process, we analyze the structural and dynamic characteristics of research interests through statistical analysis, and we also investigate on how they affect each other. As a possible application, we use observed characteristics of research interests to refine literature search on the Web, which shows that diverse user needs can be satisfied using various observations from research interests as constraints for vague queries. Such effort may 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.**

***Keywords*-research interest detection; human dynamics; retained interest; interest duration; Web search refinement**

I. 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, such as Web-based literature search systems (e.g. Google Scholar, CiteSeerX) and researchers online networks (e.g. ResearchGATE) [1]. 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 services for scientists.

In this paper, we introduce some measurement parameters to track and portray the changing process of research 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. We find that it is with some underlying principles.

Based on the acquired dynamic and structural characteristics of research interests, as an application domain of the results, we use interests (evaluated from various perspectives, such as retained interests, interest longest duration and cumulative duration) to refine literature search on the Web. A series of experiments is done based on the DBLP dataset.

II. 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.

Let $i, j \in I^+$, $y_{t(i),j}$ be the number of publications which are related to topic $t(i)$ during the time interval j.

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

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

It assumes 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.

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

Here we assume that a paper can be categorized into more than one domains 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 $avrRaI$, is the average value for all the ratio of considered research

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.246

interests in the time interval j.

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

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

$$RaI(t(i),j) = arvRaI(m,j) + \Delta RaI(t(i),j), \quad (4)$$

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

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].

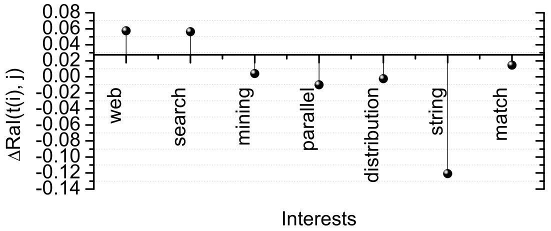

Figure 1. Relevance Ratio of Ricardo' Research Interests.

In this section, we focus on the analysis of research interests considering all the time intervals. Since they are dynamically changing, it is also important to study them chronologically.

III. 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 updated support on their research. In addition it can help to portray and support understanding on the characteristics of the dynamic process.

According to current publications, methodologies for identifying the shift of research trends can be divided into three types: the use of contents [2], the use of citations [3], and a combination of the two methods [3]. For identification of user interests on the Web, Web page content and click stream analysis has been investigated [4]. In our study, the DBLP dataset only contains author names and publication name related information (no full content or click stream data), hence we concentrate on the word-profile strategy, namely, we use word frequency to detect the dynamic change. Suppose one is interested in an area, and he/she has a steady (e.g. the same) number of publication each year, then we say the author has a steady interest in this research area. If he/she has a growing number of publication each year in an area, we say the author has a research interest growth in this research area. 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 $t(i)$ 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):

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

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 distribution.

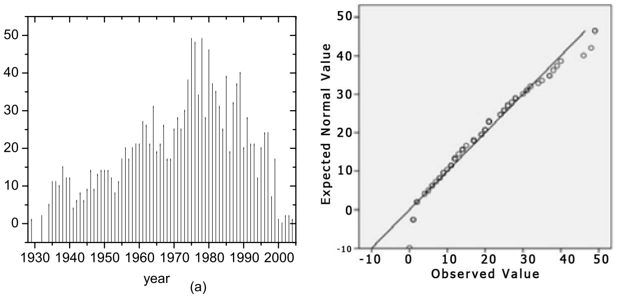

Figure 2. An analysis of degree of research interest through times. Figure 2(a) is Paul Erdos' publication distribution over years based on Erdos' publication collection (1929-1989) and MathSciNet (1990-2004). Figure 2(b) is the Q-Q diagram for Figure 2(a).

Figure 2(a) is a analysis of all the publications of a famous mathematician named "Paul Erdos". Figure 2(b) 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 years is a normal distribution.

As shown in Figure 2, no matter what kind of distribution the number of publications on a topic belongs to over years, it has ups and downs, and the period of time with a highly increasing number of publications may be belongs to a "hot" period of time. Hence, some other parameters are needed to measure this changing process of research interests.

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

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

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

[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

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

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

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 topic ($t(i)$) through the value of $DG(t(i), j)$. If $DG(t(i)) > DG(t(i'), j)$, then we say the author's research interest growth in $t(i)$ is higher than $t(i')$.

Average degree of research interest growth, denoted as $avrDG(t(i), n)$, is the average value on $DG(t(i), j)$.

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

where n is the total number of considered time intervals.

Relative degree of research interest growth, denoted as $\delta DG(t(i), k)$, is the difference from the research interest growth $DG(t(i), k)$ and 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 relevant research interest growth through an analysis of his DBLP publication data. 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.

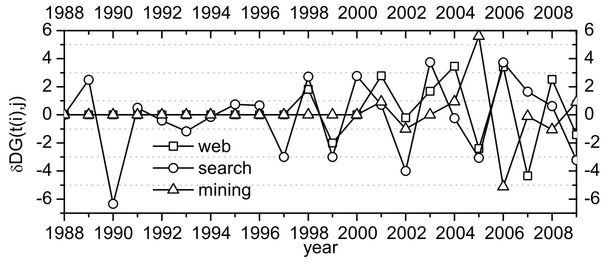

Figure 3. An analysis of Ricardo's relative degree of research interest growth $\delta DG(t(i), k)$.

Weight of a research interest, denoted as $w(t(i), j)$, is the weight of topic $t(i)$ related papers in all the papers published 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 papers related to topic $t(i)$ in the time interval j, and y_j is the total number of papers published by the author in the same specified time interval.

Suppose there are two period of time j' and j'', and 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 period of time, 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 $t(i')$.

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 having constant research interest. For example, the author has 2 published papers related to "mining" in the year 2006 and 2007, but the research interest decreased. That is because the weight of the interest got smaller.

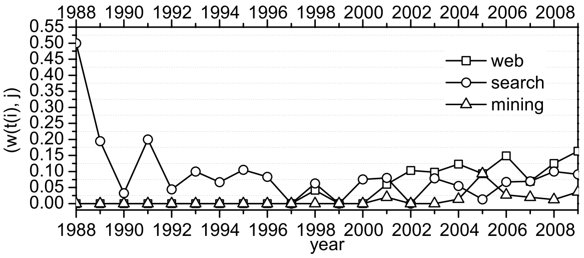

Figure 4. An analysis on the change of Ricardo's weighted interest.

The above methods and parameters only can help to identify the most recent interests based on the analysis within a time interval. Nevertheless, the impact of previous interests to the current interests has not been discussed. Here we introduce an interest model to obtain a specific user's retained interests from our previous work [5].

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 pass by if it has not appeared in some way for a long time. This phenomena is very similar to the forgetting mechanism for cognitive memory retention. Hence, we emphasize that the interest retention, which is very related to a user's current interest, can be modeled by using memory retention like functions [6]. Here we develop an interest retention model based on a power law 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 interested in topic $t(i)$ until a specified time. For each time interval j, the interest $t(i)$

might appear $y_{t(i),j}$ times, and $y_{t(i),j} \times AT_{t(i)}^{-b}$ is the total retention of an interest contributed by that time interval. According to our previous studies, the parameters satisfy $A = 0.855$ and $b = 1.295$ [5].

Here we make a comparative study on cumulative interests and retained interests. $CI(t(i), n)$ reflects a user's interest on topic $t(i)$ through all the n time intervals, which reflects the cumulative interest value. $RI(t(i), n)$ reflects a user's retained interest on topic $t(i)$ when considering the appearance of previous interests, and they focus on the interest retention on the topic in more recent years.

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 ($CI(t(i), n)$), does not always has a high retained interest value ($RI(t(i), n)$), such as "query" in the figure. In addition, although some of the interests, such as "distribution" does not have a high $CI(t(i), n)$ value, they may have very high $RI(t(i), n)$ values since they may be currently, at least most recently interesting to a user.

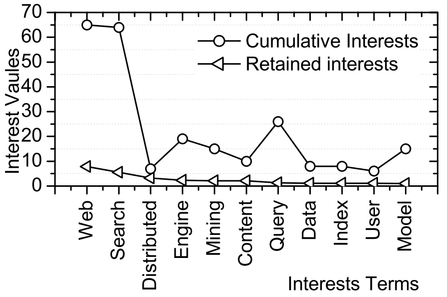

Figure 5. A comparative study on the cumulative interests and retained interests of the author "Ricardo A. Baeza-Yates" based on the author's publication list from to 2009.

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 [3].

IV. Building and Analyzing the Structure of Research Interests

In this section, we firstly examine the structure of research interests from the network perspective. Then we investigate on the dynamics of these structures in a chronological order.

A. Constructing the structure of research interests

All the research interests can be connected together to form a networked structure. Figure 6 provides some examples of research interests networks. It shows how interests (here we evaluate the hot topics by their values of cumulative interest $CI(t(i), n)$) shift in a timely manner (we choose the top 8 ranked single-word topics from the year 1991, 1997, 2003, 2009). Since we investigate the problem in a network setting, the selected interests are pivotal nodes in the networks, hence the shift of them shows the major dynamic changing process on the shift of research interests. Some interesting phenomena have been observed:

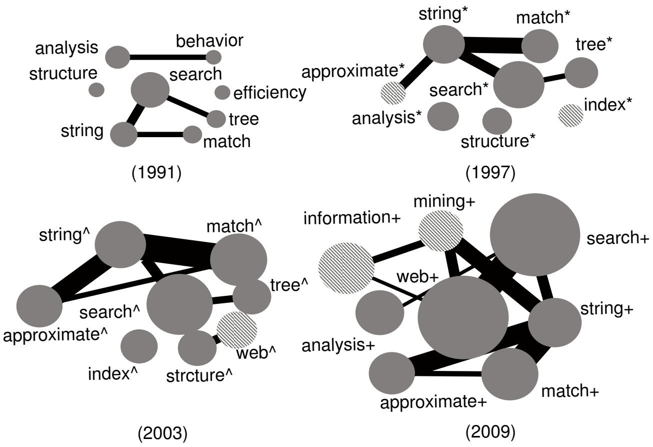

Figure 6. Ricardos research interest dynamic evolution network from 1991 to 2009. (Based on DBLP publication list, with 232 papers involved). The network is a graph with weighted edges and weighted vertices.

(1) In the interests networks, pivotal nodes are dynamically changing all the time. Some of them are growing larger (e.g. search), which may be due to a growing interest in the topics, and some of them disappear from the top 8 pivotal nodes (e.g. tree, behavior), which may be due to the lost of interests. Meanwhile, some new interests emerged (the ones that are marked with decorative patterns, e.g. web).

(2) Some top research interests remain active in the interests networks (e.g. search, analysis, match).

(3) Main research interests are closely related to each other, which made the degree of separation around 2-3. This phenomenon indicates that an author's research interests are not isolated, instead, they are highly relevant.

(4) The width of the link shows the degree of connections for two single-word terms (if both of them appear in the same paper, then one degree of connection is added to them). If the author has interest in working on the synergy of two related topics, then connections between them will grow stronger as time goes by. The figure shows that relations among research interests varies chronologically (e.g. the connections between "Web" and "search").

B. Analyzing the Structure of Research Interests

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

Degree characteristics of research interests :

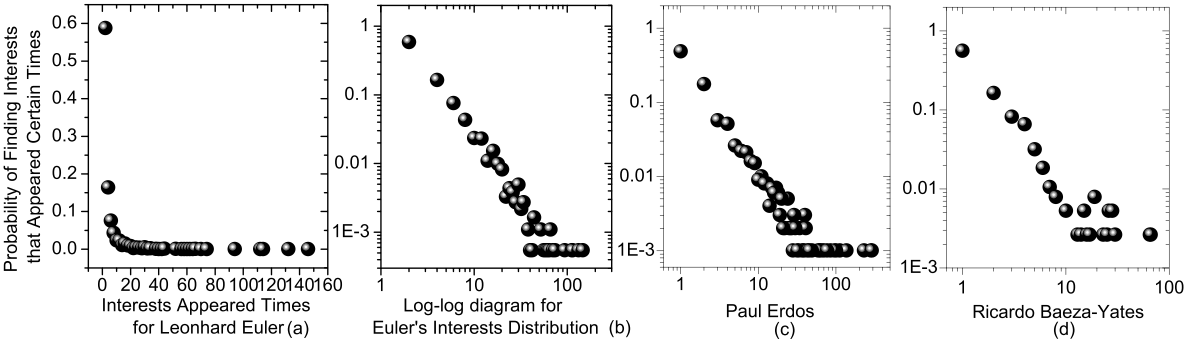

Figure 7. Power-law distribution on weights of research interests for Leonhard Euler (Publication list is from Euler's Archive), Paul Erdos (publication list is from Erdos' publication collection and MathSciNet), and Ricardo Baeza-Yates (publication list is from DBLP)

Concerning the weight of research interest for each single-word topic term, only a few of them are with high weights of research interest, and most of them are with low weights.

We examined three authors' degree destribution of research interests. Figure 7 shows that research Interests for these authors follow power-law distribution. The slopes are -1.62±0.15(Euler), -1.15±0.07(Erdos), and -1.33±0.14(Ricardo) respectively. It shows that in scientific research, we may approximately consider the slope values are close to each other for different authors, although not that close as people observed in other human activities, such as mail correspondence (with the slope value 1.5 [7]).

The structural characteristics of research interests shows that there are always some major interests in the network. By preferential attachment theory [8], we can conclude that new interests in the network prefer to be connected with the existing major interests. Hence, main research interests are of vital importance to researchers' future interests [8].

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 [9]. Recent findings emphasize that consider from the time perspective, many human activities (e.g. email and short message sending, online clicking of web pages, making calls, financial commerce, etc.) follow power-law distribution [9], [10], [11], [12]. The results indicate that there might be deeper underlying principles for human activities.

Scientific research is a typical human activity, and the process on the shift of research interest is in a timely manner. To the best of our knowledge, there is few study on timing statistical characteristics on the shift of research interests.

A single research interest's distribution over years may not follow the same type of probability distribution. 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 have a 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 back to another burst, may have power-law distribution. Nevertheless, when we put all the interests in a box and investigate them, some interesting phenomena can be observed.

The process on the shift of research interest is to some extend 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 enables us to investigate on the statistical characteristics of the 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 at 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 $t(i)$:

$$ILD(t(i)) = \max(ID(t(i))_n), \tag{13}$$

where $n \in I^+$, $ID(t(i))_n$ is the interest duration when $t(i)$ discretely appears (the time interval of the appeared interest is not directly continuous with the one of the previous appeared interest) for the nth 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:

$$ICD(t(i)) = \sum_{n=1}^{n'}(ID(t(i))_n), \tag{14}$$

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

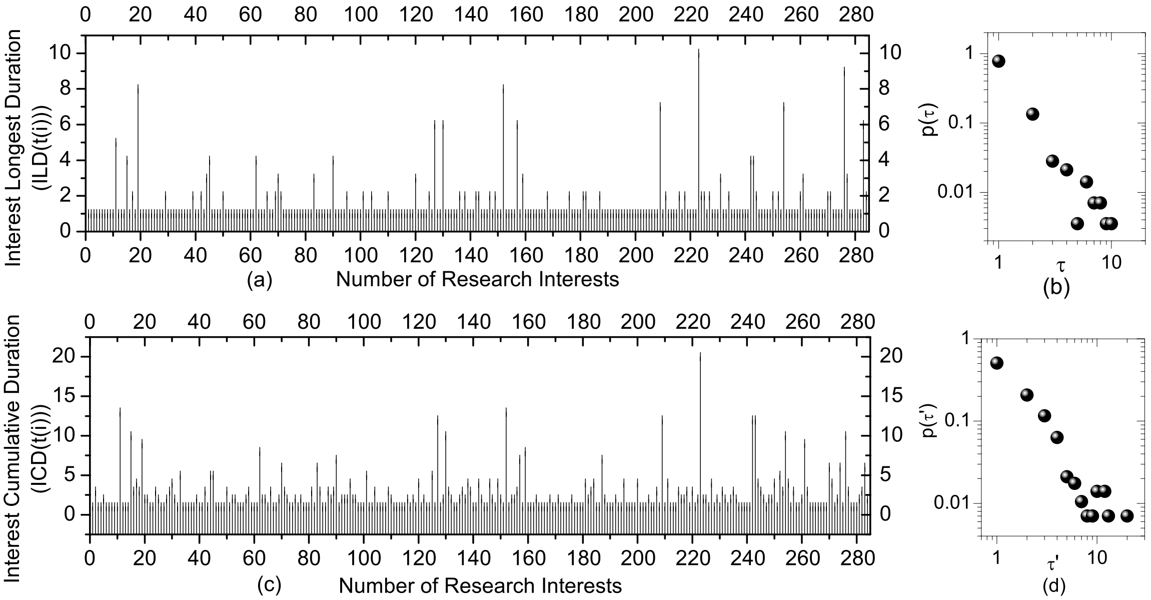

Figure 8. Ricardo's research interest lasting time and appear time statistics.

Figure 8(a) 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 8(b) 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:

$$P(\tau) \approx \tau^{-\alpha}, \tag{15}$$

where $\alpha \simeq 1.64$ (the solid line in the log-log plot has slope -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 8(c) is an analysis of Ricardo's $ICD(t(i))$. We can observe similar phenomenon as in figure 8(a), that there are some large spikes in the plot, corresponding to very long $ICD(t(i))$ for some research interests. As shown in figure 8(d), the statistical distribution on the value of $ICD(t(i))$ can be best approximated as:

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

where $\alpha' \simeq 2.30$ (the solid line in the log-log plot has slope -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 8(a) and Figure 8(c) shows that the $ILD(t(i))$ and $ICD(t(i))$ for a $t(i)$ do not always consistent (the x-axis of these two figures share the same corresponding interests), namely, one research interest may have appeared in many years, hence have relatively longer $ICD(t(i))$, but has a relatively short $ILD(t(i))$, 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 on the interest longest duration 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 seems 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 figure 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 a 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 a relatively small number of appearance times.

We 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 the e-foaf:interest vocabulary [13] is used to describe them in an RDF file [2].

V. Search Refinement by Research Interests from Different Perspectives

From the network theory perspective, the process of investigating unexplored topics can be considered as adding new nodes to the interests network. By the phenomena of preferential attachment which has been briefly discussed in Section IV-A, we can predict that unexplored topics are likely to be connected with big 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 [14]. Hence, research interests can be considered as a context for literature search on the Web. When the query is vague/incomplete, research interests can serve as constraints that can be used to refine these queries. Research interests can be evaluated from various perspectives and each perspective reflects one unique characteristics of them. As an illustrative example, based on the study above, we examine the research interests from 3 perspectives introduced in Section II and Section IV-B.

Table I
Top 9 interests with the biggest retained interest (RI) values, with the biggest interest longest duration (ILD) or interest cumulative duration(ICD) (User name: Ricardo A. Baeza-Yates)

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
data	1.13	information	6	structure	12
index	1.09	mining	6	index	12

Table I is a comparative study of an author's top 9 interests with the biggest interest retention 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 obtained interests through these perspectives. Table II 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 this three perspectives, different search results are selected out and provided to users to meet their diverse needs (In this partial list of results, literatures with the query keywords and constraints from research interests are selected out and ranked to the front. As an illustrative example, in each list, our system shows the first search results that are obtained according to constraints from each of the research interests).

Based on the above study, we developed a literature search system with mentioned search refinement functionalities using various research interests models based on the DBLP dataset. The assumption is that the users are willing to log on the system with their real names and they need to have some publications that are recorded in the DBLP dataset. Through this system and above studies, one can get a preliminary idea on how the research interests evaluated from various perspectives serve as an environmental factor that affect the search refinement process and help the researchers get more relevant search results for further investigations.

VI. Conclusion

This paper concentrates on the study of explicit research interests that appeared in authors' previous publications and uses them as contextual foundations for Web search refinement. The dynamic and structural characteristics of research interests are investigated. From the perspective of dynamics, in this paper, we provide some preliminary methods for tracking the dynamic changing process of research interests. From the perspective of structures, by utilizing network theories, we investigate the statistical distribution on the structures and evolution process of interests networks and provide some basic understanding on the evolution characteristics of the interests networks.

In this paper, 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. After finding these characteristics, we are going to consider multiple word terms. For scientists, their research interests is 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.

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, we provide some illustrative examples on how to refine the search process using acquired interests from different perspectives. This can be considered as some efforts towards user centric knowledge retrieval [15].

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

Table II
SEARCH REFINEMENT USING THE TOP 9 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 9 interests that have the biggest retained interest values
	Web, Search, Distributed, Engine, Mining, Content, Query, data, index
	* 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. * ......
List 2 :	with the top 9 interests that have the biggest interest longest duration
	search, web, text, match, approximate, retrieval, query, information, mining
	* 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. ......
List 3 :	with the top 9 interests that have the biggest interest cumulative duration
	search, retrieval, algorithm, text, match, query, string, structure, index
	* 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. ......

ACKNOWLEDGEMENT

This study is supported by the European Commission under the 7th framework programme, Large Knowledge Collider (LarKC) Project (FP7-215535). We would like to thank Yiyu Yao, Rui Guo, Chao Gao, Jian Yang on their constructive discussions.

REFERENCES

[1] B. Shneiderman. Science 2.0. *Science*, 319:1349–1350, March 7 2008.

[2] C. Erten, P.J. Harding, S.G. Kobourov, K. Wampler, and G. Yee. Exploring the computing literature using temporal graph visualization. In *Proceedings of the 2004 SPIE Conference on Visualization and Data Analysis*, volume 5295, pages 45–56, 2004.

[3] C.M. Chen. 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, 2006.

[4] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *Proceedings of the 2006 International World Wide Web Conference*, May 2006.

[5] Y. Zeng, Y.Y. Yao, and N. Zhong. Dblp-sse: A dblp search support engine. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 626–630, September 2009.

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

[7] J.G. Oliveira and A.L. Barabási. Darwin and einstein correspondence patterns. *Nature*, 437(1251), 2005.

[8] A.L. Barabási. *Linked: How everything is connected to everything else and what it means for science, business and everyday life*. Perseus Publishing, 1 edition, 2002.

[9] A.L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

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

[11] X.P. Han, T. Zhou, and B.H. Wang. Modeling human dynamics with adaptive interest. *New Journal of Physics*, 10(073010), 2008.

[12] J. Masoliver, M. Montero, and G.H. Weiss. Continuous-time random-walk model for financial distributions. *Physical Review E*, 67(021112), 2003.

[13] Y. Zeng, Y. Wang, Z.S. Huang, D. Damljanovic, N. Zhong, and C. Wang. User interests: Its definition, vocabulary, and utilization in unifying search and reasoning. In *Proceedings of the 2010 International Conference on Active Media Technology*, August 28-30 2010.

[14] J.D. Bransford, A.L. Brown, and R.R Cocking. *How People Learn: Brain, Mind, Experience, and School*. National Academy Press, 2000.

[15] Y.Y. Yao, Y. Zeng, N. Zhong, and X.J. Huang. Knowledge retrieval (kr). In *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 729–735, 2007.