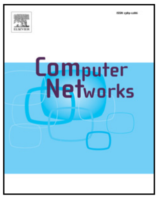
Computer Networks 113 (2017) 205–217

Contents lists available at ScienceDirect

Computer Networks

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



On the most representative summaries of network user activities

Joshua Stein[a], Han Hee Song[b], Mario Baldi[b,c,*], Jun Li[a]

[a] University of Oregon, 120 Deschutes Hall, Eugene, Oregon, USA
[b] Cisco Systems, 255 W Tasman Dr., San Jose', California, USA
[c] Politecnico di Torino, C.so Duca degli Abruzzi 24, Turin, Italy

ARTICLE INFO

Article history:
Received 21 May 2016
Revised 15 November 2016
Accepted 23 December 2016
Available online 30 December 2016

Keywords:
Network user activities
User interest discovery
Interest summarization

ABSTRACT

A summary of a user's Internet activities, such as web visitations, can provide information that closely reflects their interests and preferences. However, automating the summarization process is not trivial as the summary should strike a good balance between generality and specificity, while there is no gold standard for doing so.

In our approach to summarizing user information, dubbed SUM, we develop two scoring mechanisms that cooperatively optimize for polarizing criteria. After mapping user activity information onto a category tree, the scoring mechanisms highlight the most representative tree node (or summary); the node provides an aggregated view of the activities most characteristic of the user. We evaluate our approach by using web activity on the network of a large Cellular Service Provider and summarizing it to devise interests of individual users as well as groups. We compare SUM against an algorithm that discovers Hierarchical Heavy Hitter and show that SUM uncovers previously unknown information about users.

1. Introduction

As people are heavily connected to each other through the Internet more than ever before, they communicate a substantial amount of information about themselves. Such information is embedded in their network traffic, such as website visitations, data exchanged using different (mobile) applications, and GPS coordinates sent from their mobile devices. This data provides an opportunity to discover rich information about the users that can be of very high value to communication service providers, hence ultimately to users themselves. It is well known that the many services available free of charge on the Internet, such as search engines, social media, news, e-mail accounts, are offered in exchange for the user authorization to use the information she shares for commercial purposes (most commonly advertisement). However, one of the services that we are offered no other option other than paying for it, is access to the Internet. In fact, connectivity providers have no easy way of monetizing on data collected about their customers. While they are in principle in the ideal position as they can potentially access information shared through all of the specific services, such a vast amount of data is difficult to consume because it is composed of an extremely large number of very detailed items. A way to harness such a valuable resource would enable connectivity provider to give their customers the option of receiving fixed and mobile Internet connectivity free of charge in exchange for the consent for the provider to tap into the information that users exchange.

Summarization is the key to make such wealth of information manageable and practically usable, thus giving its users an opportunity to benefit from its value. The difficulty in summarizing the diverse sets of information is that there is no golden rule to objectively assess weights of different activities, i.e., how representative they are of users and their interests.

Example. *A content provider (CP) may be interested in understanding its users' cyber activities. When a user, Alice, shops for skates from an Internet shopping site, the CP may consider her interested in shopping, sports, or both. When new logs of Alice visiting a webpage of a ski resort come in, the CP may consider her to be into 'winter sports'. If new data reveals that she also frequents score boards of baseball, basketball, etc. on ESPN, the CP may consider Alice's interests to be simply in 'sports'.*

As depicted in the above example, summarizing a user's interests is far from being trivial. Some inputs may broaden the scope of a user's interests, some may narrow it down. Hence, a systematic method that strikes a good balance between generality and specificity is needed.

A natural approach to this problem is to depict it as a graph of interests and determine which interest is the most significant. The

[*] Corresponding author.
E-mail addresses: jgs@cs.uoregon.edu (J. Stein), hanhsong@cisco.com (H.H. Song), mario.baldi@polito.it (M. Baldi), lijun@cs.uoregon.edu (J. Li).

http://dx.doi.org/10.1016/j.comnet.2016.12.014

computation of significance falls under a class of problems known as graph centrality, with PageRank [1] being one method to solve it. However, methods such as PageRank are insufficient for computing the most significant node in a *structured graph* such as one where user interests are represented as a tree.

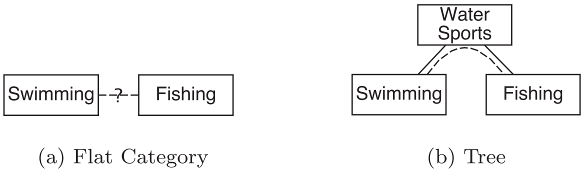
Research on finding Hierarchical Heavy Hitters (HHHs) addresses specifically the case of tree-structured data [2–4]. Methods for the identification of HHHs are particularly effective in summarizing activities in IP prefix-based *trie* structures. However, as their target applications are limited to network topologies where associations among tree nodes are strictly enforced by the IP addressing, these algorithms cannot be directly applied to our context where nodes are associated through softer, *semantic* similarities. Therefore, a more general approach is necessary for summarizing data categorized in ways less structured than IP prefixes.

In this paper, we develop an approach, referred to as **SUM (Summarizer for User inforMation)**, that flexibly summarizes users' network activities into information about the users and their interests. To allow fair comparison among various network activities a user conducts (such as web browsing, usage of mobile apps, sharing of information), SUM maps them onto a single category hierarchy. Once the user activities are standardized, it then searches for the most representative summary of the activities within the hierarchy. For this, we develop two scoring methods based on Graph Centrality [5] — *choice score* and *stop score* — to perform a search on the category hierarchy. Beginning from the root of the tree, for each node, we assign a choice score which represents the preference of a direction, i.e., a child node, to traverse further. At the same time, we assign a stop score which is used in determining a sweet spot between choosing a general vs. specific (i.e., deeper in the hierarchy) node on the branch the choice score chooses. Traversing the tree based on the two scoring mechanisms, SUM identifies a tree node that best represents the user's activity.

Challenges. The approach we developed for summarizing user information heavily depends on the structure of *data categorization*, which, being built by humans (*i.e.*, domain experts), is prone to be imperfect. For example, the tree might contain *inaccurate semantic structures* whereby children of nodes in the ontology tree may not be completely covered by the semantics of their parent. We do not place restrictions on the *topological properties of ontology trees, i.e.*, the number of children or ancestors a node may have. In addition, due to the subjectivity in determining what is a good summary, we *lack ground truth* to evaluate how well SUM summarizes user information.

Addressing the above challenges, we run SUM on a dataset of web visitations from a large CSP (Cellular Service Provider) in North America and map them onto an ontology with 65,000 user activity categories. We analyzed the dataset from the perspective of multiple different applications for our approach. To the best of our knowledge, this is the first work that systematically summarizes network user activities in a large scale. Our approach makes the following specific contributions:

- We design and implement the SUM algorithm that flexibly chooses a representative summary which has an appropriate balance between being general and specific. Hinged on the concepts of graph centrality, for a user (or a group of users), our algorithm determines the most representative yet specific summary from a pool of hierarchically classified activities. Our algorithm allows analysts to easily tune the summary to be between general and specific, depending on the application.
- We evaluate SUM using web browsing history collected from a large CSP covering 4 million web visits from 150,000 Internet users for five days. We demonstrate how SUM is able to summarize the web browsing interests of individuals as well as groups of users. The summaries that are produced by SUM have high stability under varying amounts of interest data differing by 1.37 categories on average, and possess a common depth (i.e., specificity) of 2.42 on average. Furthermore, the summaries produced by SUM are relevant even in the presence of skewed interests.
- We perform a case study with SUM and HHH on a collection of web browsing activity generated by a large number of Internet users. SUM and HHH agreed on many of their most specific categories, but SUM proved to have more descriptive and representative summaries than HHH. SUM revealed properties of the users' Internet usage, particularly honing in on their preferred search applications. Beyond search, users were found to have a particular affinity for visiting footwear shopping sites.



(a) Flat Category (b) Tree

Fig. 1. Expressiveness of categorical semantics. Dashed line indicates an implicit relationship.

2. Background

In this section we define the components necessary for summarization in addition to the properties a summary should exhibit. The summarization process requires data, to which we add annotations in the form of the category of an ontology the data may fall under. In this work the semantics of the annotation is restricted to be defined by an ontology composed of categories the data may fall under. The summary that results from the data must be general enough to capture a good fraction of the data, while still being specific enough such that no significant information is lost. In other words, building a summary takes up the challenge of striking a balance between the breadth of data captured, as well as its depth.

2.1. Categorization ontology

The ontology the summarization process operates on has relationships between categories. The number of relationships for a category is not restricted, but the topology of the ontology must be a *tree*. The motivation for using a tree over flat categorizations or a general graph is due to the explicit relationship between categories and expressiveness of the topology, respectively.

Tree-based vs. flat categorization: Flat categorization, or keyword categorization, explicitly states the category that data falls under. The limitation of utilizing a single keyword is that keywords do not relate to one another, which would enable us to derive strong connections between data items. If a strong connection were present between single keywords then we would be able to construct a structured categorization, which we will cover more generally below. For instance, categories such as "swimming" and "fishing" may be used for labeling data but they lack any indication of relationships between them (*i.e.*, Fig. 1(a)). Conversely, in a tree topology having a single ancestor, such as "water sports", expresses implicitly (and compactly) a relationship between the two nodes (*i.e.*, Fig. 1(b)).

Tree-based vs. general graph-based categorization: A general graph is able to accurately fulfill the expressiveness of an ontology tree as well as more complex relationships. A category could be reachable through multiple paths from the most general category (i.e., the one at the root of the tree) and thus obtain different meanings depending on the path. An example of this would be