摘　要

随着网络技术的发展，人类已经进入互联网时代，越来越多的用户选择通过互联网获取信息。但是，由于网络数据量的急速增长，使得用户往往迷失在信息海洋中，难以快速寻找出感兴趣的内容。因此，如果能根据不同用户的需要，快速准确地检测出网络上存在的话题可以大大减少信息搜索的时间，从而提高用户体验。在本文，基于网络话题和网络数据的特点，我们提出了三种方法，以帮助用户快速寻找到关注的话题。这三种方法分别是网络话题检测、网络话题优化和网络话题检索。通过在 MCG-WEBV 和 YKS 两个数据集上进行实验，证实了这三个算法的有效性。

首先，本文研究了网络话题的多粒度性，多粒度性是由网络话题在不同相似度层上的演化造成的。因此，本文提出了基于相似度流的多粒度网络话题检测算法。首先，我们用相似度图表示网络数据之间的关系；然后，在一系列门限下对相似度图进行截断，并在每个截断图中寻找所有的极大团，所有截断图中的所有极大团就是数据集的候选话题；之后，用候选话题重构原始相似度图，计算每个候选话题的重构系数；最后，根据重构系数对候选话题进行排序，并返回有序话题序列，排序靠前的话题是网络上存在的话题，排序靠后的话题是其多粒度表现。此外，现有评测方法只考虑算法检测话题的精度，对此本文提出了一种新的评价准则，既能计算话题检测的精度，也同时能够衡量算法的检测代价。

其次，本文研究网络话题的优化。在目前所有的算法中，检测到的网络话题均不完善——有些属于话题的网页并未被检测到，而检测到的话题内的某些网页却并不属于该话题。首先构建异质图来表示待优化话题、网页和单词之间的关系，然后在异质图上引入随机游走算法以计算待优化话题在所有网页上的概率分布，并根据概率值进行话题的优化。与此同时，针对话题大小难以确定的问题，提出了一种自适应的确定方法。

最后，本文提出一个交互式多模态网络话题检索系统。话题具有较强的主观性，不同的用户对话题具有不同的认识，为了寻找到真正感兴趣的话题，需要用户提供描述信息。在系统中，同样使用异质图表示网页、单词和图像之间的关系，然后在异质图上使用重启动随机游走算法计算用户查询与各个网页之

间的关系，并根据这些关系确定话题检索的结果。系统提供了两种方法确定检索话题的大小，一种方法是通过用户指定，另一种方法是系统自适应确定。在系统中，用户可以对话题进行缩放，从而快速定位到感兴趣的内容。如果用户对系统检索到的话题不满意，可以提供更多的描述信息，通过人机交互优化检索结果。

关键词：网络话题，检测，优化，检索，多粒度

ABSTRACT

Fei Jia (Computer Technology)
Directed By Associate Professor Yugui Liu

We are now in Internet age due to the rapid development of web technology, as more and more people access information from Internet. However, ever-increasing amounts of data on the web make it hard for users to find the content they are really interested in. Therefore, if an effective and efficient method to find topics on the web can be provided for different people, it naturally not only reduces the searching time but also enhances user experience. In this thesis, based on the character of web topics and web data, we propose three methods, web topic detection, web topic boosting and web topic search, to help users find their concerned topics rapidly. Experiments on two public data sets, i.e., MCG-WEBV and YKS, demonstrate the effectiveness of these algorithms.

Firstly, multi-granularity of topics is researched in this thesis. Multi-granularity is caused by topic evolution across different similarity levels. Therefore, we come up with a multi-granularity topic detection algorithm that is based on similarity cascades. A similarity graph is used to represent the relationships among web data. We truncate the similarity graph at a series of thresholds, and find all maximal cliques as topic candidates from these truncated graphs. Then, all topic candidates are used to reconstruct the original similarity graph and the reconstruction coefficients are calculated. These topic candidates are sorted by their reconstruction coefficients and the sorted topics are returned to users. Topics at the top of the list are treated as detected topics and others are multi-granularities of these detected topics. Besides, since the existing evaluation criterion only considers the accuracy of topic detection algorithms, a new criterion is proposed in this thesis, which not only calculate the detection accuracy, but also evaluate the false positives of algorithms.

Secondly, a web topic boosting problem is proposed in this thesis. Topics detected by all existing algorithms are not perfect, that is, there are some web pages belong to a topic but not be detected, and some web pages in a topic but don’t belong to it. We first construct a heterogeneous graph to represent the relationships among unboosted topics, web pages and words. Then random walk on the heterogeneous graph is adopted to calculate the probability distribution of an

unboosted topic on all web pages, and the probabilities are used to boost the topic. Moreover, due to the difficulty of determining topic size, this thesis proposes a self-adapting method.

Thirdly, this thesis proposes a multi-modality topic search system. Web topic is subjective, requiring the understanding of a topic from various viewpoints. In order to search interesting topics, we need some extra description information provided by users. In this system, a heterogeneous graph is constructed to represent the relationships among web pages, words and images. Then random walk with restart algorithm is used to measure the relationships between user queries and all web pages. Our system provides two ways to determine topic size, one is user assignment, and another one is self-adapting determination. In this system, users can zoom a topic to quickly locate concerned content. If users are not satisfied with the searched topic, they can refine it by user interaction.

Key Words: web topic, detection, boosting, search, multi-granularity