Public Abstract
First Name:Suman
Middle Name:
Last Name:Deb Roy
Adviser's First Name:Wenjun
Adviser's Last Name:Zeng
Co-Adviser's First Name:
Co-Adviser's Last Name:
Graduation Term:FS 2013
Department:Computer Science
Degree:PhD
Title:ON CROSS-DOMAIN SOCIAL SEMANTIC LEARNING

Approximately 2.4 billion people are now connected to the Internet, generating massive amounts of data through laptops, mobile phones, sensors and other electronic devices or gadgets. Not surprisingly then, ninety percent of the world's digital data was created in the last two years. This massive explosion of data provides tremendous opportunity to study, model and improve conceptual and physical systems from which the data is produced. It also permits scientists to test pre-existing hypotheses in various fields with large scale experimental evidence. Thus, developing computational algorithms that automatically explores this data is the holy grail of the current generation of computer scientists.

Making sense of this data algorithmically can be a complex process, specifically due to two reasons. Firstly, the data is generated by different devices, capturing different aspects of information and resides in different web resources/ platforms on the Internet. Therefore, even if two pieces of data bear singular conceptual similarity, their generation, format and domain of existence on the web can make them seem considerably dissimilar. Secondly, since humans are social creatures, the data often possesses inherent but murky correlations, primarily caused by the causal nature of direct or indirect social interactions. This drastically alters what algorithms must now achieve, necessitating intelligent comprehension of the underlying social nature and semantic contexts within the disparate domain data and a quantifiable way of transferring knowledge gained from one domain to another. Finally, the data is often encountered as a stream and not as static pages on the Internet. Therefore, we must learn, and re-learn as the stream propagates.

The main objective of this dissertation is to develop learning algorithms that can identify specific patterns in one domain of data which can consequently augment predictive performance in another domain. The research explores existence of specific data domains which can function in synergy with another and more importantly, proposes models to quantify the synergetic information transfer among such domains. We include large-scale data from various domains in our study: social media data from Twitter, multimedia video data from YouTube, video search query data from Bing Videos, Natural Language search queries from the web, Internet resources in form of web logs (blogs) and spatio-temporal social trends from Twitter.

Our work presents a series of solutions to address the key challenges in cross-domain learning, particularly in the field of social and semantic data. We propose the concept of bridging media from disparate sources by building a common latent topic space, which represents one of the first attempts toward answering sociological problems using cross-domain (social) media. This allows information transfer between social and non-social domains, fostering real-time socially relevant applications. We also engineer a concept network from the semantic web, called semNet, that can assist in identifying concept relations and modeling information granularity for robust natural language search. Further, by studying spatio-temporal patterns in this data, we can discover categorical concepts that stimulate collective attention within user groups.

Using these various disparate data from different domains, my dissertation aims to assert that intelligent learning is a mixture of two parts: combinatorial knowledge representation from diverse data, and transferring the gained knowledge appropriately to tackle a new task which could not be solved elegantly without the synergy. In summary, this work demonstrates that traditional learning models for classification, prediction and recommendation (such as Support Vector Machines, Latent Dirichlet Allocation, Genetic Algorithms, Conditional Random Fields, Decision Trees, Path Analysis, Probabilistic Automata) can be boosted by algorithmically transferring related social and semantic data from cross-domains.