Abstract

For any work of literature, a fundamental issue is to identify the individual(s) who wrote it, and conversely, to identify all of the works that belong to a given individual or to identify the individual who writes many papers on the same topic or to identify the topics name that an author works on. Information extraction techniques (such as Author Name and Topic Recognition) have long been used to extract useful pieces of information from text. The types of information to be extracted are generally fixed and well defined. However in some cases, the user goal is more abstract and information types cannot be narrowly defined. For example, a reader of online user reviews typically has the goal of making a good choice and is interested to learn about the different aspects of a topic and author relation (e.g., famous author of a topic, author’s papers with his research field). Some of these aspects may be known by the reader and some others may need to be discovered from the inherent text structure in a large collection. Even for the known aspects (such as “author name” and “topic”), the challenge is to recognize various hidden aspects like number of papers written by an author, his research field, popularity of an author.
In this paper, we will develop content modeling Paradigm to extract the relationship between the author, document, topic and Words as topics with identifiable word distributions across documents of various authors. We review several probabilistic graphical models (such as Latent Dirichlet Allocation) and propose a new model called content modeling paradigm which is based on frequency of the words within the document.

**Reference**

  - Integrating topics and Syntax. In Advances in NIPS 17.
  - Next-generation Intrusion Detection Expert System by Debra Anderson Thane Frivold
  Alfonso Valdes Computer Science Laboratory 1995

Index Terms

Computer Science

Soft Computing
Key words

supervised paradigm

unsupervised paradigm

ATP Model

TAP Model