微博标签自动生成策略调研

标签内容：

兴趣爱好、职业领域、自我性格、形象、自我描述

生成策略：

1. 基于关键词的生成方法

主要包含TFIDF与TextRank，TextRank效果较好：基于词语的共现关系，构建词语网络，抽取较为重要的词即TextRank值较高的词作为标签。关键词再拓展：将相邻的词语进行组合。

2. 基于社会支持网络的生成方法

作者认为，不同用户可能因为对某个特定的用户的不同方面感兴趣而关注他，因此通过建立用户的粉丝发布的微博主题模型挖掘标签。

3. 基于聚类分析的生成方法

文本预处理（取TextRank权重前200）获得名词作为候选关键词，计算任意名词对间的相似度，进行聚类分析，取top10聚类簇。

4. 基于分类的生成方法

将用户感兴趣的若干个类别作为标签，人工构建目标分类体系及微博训练语料，使用SVM识别出用户感兴趣的类别作为标签。

5. 基于百度百科的生成方法

利用百度百科具有三层分类信息的词条资源，识别出用户关注的类别作为标签。

评价方法：

人工评价，根据一致性评定筛选有效值。