数据库系统概论新技术篇

社交网络大数据分析
--社交媒体初探（2）

赵鑫

中国人民大学信息学院

2017年5月

讲述提纲

社交媒体概述

常用数据处理技术

典型任务与解决方法

研究方向展望

小结

典型任务与解决方法

数据质量清洗

用户舆情分析

用户画像构建

用户兴趣学习

典型任务与解决方法（续）

- 数据质量清洗
  - 数据标准化
  - 谣言检测
  - 水军检测

数据规整:

信息表述中存在歧义、多义以及新词。如，郭敬明→小四、买苹果、word哥

常用技术：

多义/歧义：词义消歧、实体链指（知识图谱）

新词：大规模语料学习规律

典型任务与解决方法（续）

- 数据质量清洗
  - 数据标准化
  - 谣言检测
  - 水军检测

谣言检测:

社交媒体中经常会出现很多虚假新闻，如“某名人被去世”

常用技术：

以分类器为主要模型，融入多种信息特征，包括：
1、发布者身份
2、多信息源
3、新闻反馈（用户评论）

典型任务与解决方法（续）

- 数据质量清洗
  - 数据标准化
  - 谣言检测
  - 水军检测

水军检测:

虚假信息的发布用户，例如电商平台的虚假评论发布者

常用技术：

以分类器为主要模型，融入多种信息特征，包括：
1、发布者身份
2、发布者行为特征
3、发布文本特征

典型任务与解决方法

数据质量清洗

用户舆情分析

用户画像构建

用户兴趣学习

典型任务与解决方法（续）

- 用户舆情分析
  - 情感抽取
  - 情感预测
  - 摘要生成
  - 趋势预测

情感抽取:

有效抽取情感与属性（主题）词汇，
例如 The phone is **good** to **use**.

常用技术：
采用序列标注模型，融入多种信息特征，包括：
1、词、词性特征
2、词向量特征
3、词组合特征

典型任务与解决方法（续）

- 用户舆情分析
  - 情感抽取
  - 情感预测
  - 摘要生成
  - 趋势预测

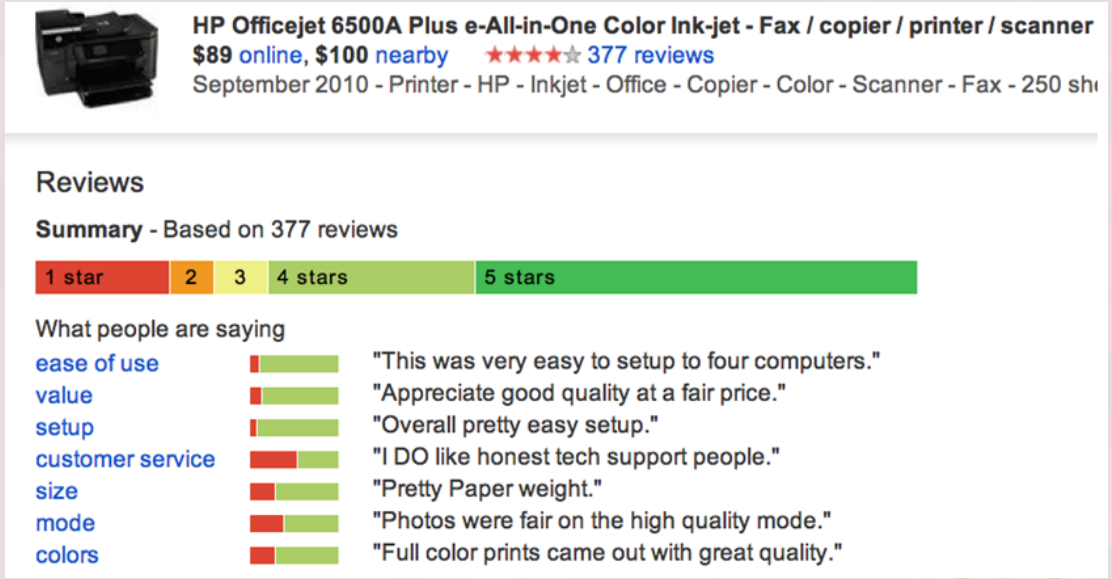
评分预测（褒贬分析）：

针对用户文本，预测其舆情取向或者评分。如，预测产品评分。

常用技术：
主要采用回归或者分类模型，融入多种信息特征，包括：
1、词、词性特征
2、词向量特征
3、词组合特征
4、复杂语义特征（深度学习）

典型任务与解决方法（续）

- 用户舆情分析
  - 情感抽取
  - 情感预测
  - 摘要生成
  - 趋势预测



典型任务与解决方法（续）

- 用户舆情分析
  - 情感抽取
  - 舆情预测
  - 摘要生成
  - 趋势预测

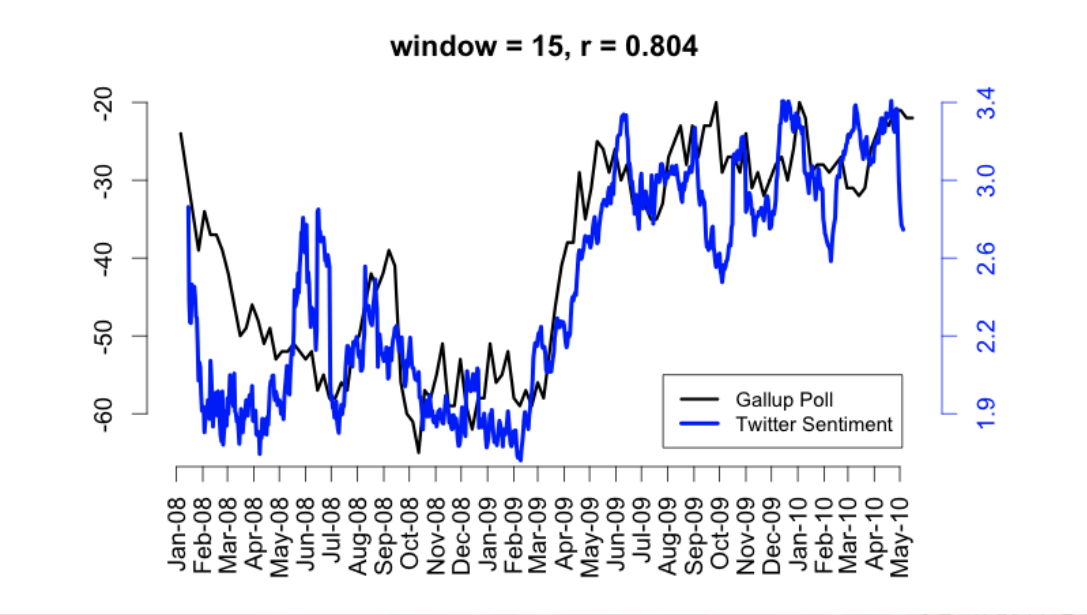
摘要生成（褒贬分析）：

针对特定产品或者话题的用户文本，
生成属性或者话题的舆情摘要。

常用技术：
概率主题模型

典型任务与解决方法（续）

- 用户舆情分析
  - 情感抽取
  - 舆情预测
  - 摘要生成
  - 趋势预测



Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010

典型任务与解决方法（续）

- 用户舆情分析
  - 情感抽取
  - 舆情预测
  - 摘要生成
  - 趋势预测

趋势预测:

利用社交媒体中用户的舆情趋势，来预测未来某种数据指标的变化趋势，如股票、产品销量。

常用技术：
首先，要发现并确认关联模式。
其次，建立预测模型（如时间序列）。
最后，融入舆情信息。

小结

了解数据质量清洗以及用户舆情分析两个任务的基本技术手段。