南京大学
NANJING UNIVERSITY

一种基于多维时序分析的音乐推荐系统及其实现

姓名：王守涛

导师：徐锋 教授

时间：2014.5.8

汇报提纲

- 背景介绍
- 基于多维时序分析的音乐推荐
- 基于用户三期行为的综合音乐推荐
- 系统实现
- 总结与展望

背景介绍

- **问题背景**[Song+@CMMR'12]
  - 信息过载
  - 选择悖论
- **音乐特点**[Ocelma+@RecSys'11]
  - 时间短、消费代价低
  - 划分标准不一、多重属性
  - 与次序相关、上下文相关

背景介绍

- **问题描述**[Park+@CNSI'11]
  - 分析用户所收听的歌曲，预测其可能收听的下一首歌曲
  - 为用户推荐一个歌曲列表，使之尽可能地符合用户即时需求

$$U = \{u_1, u_2, \ldots, u_v\}$$

$$S = \{s_1, s_2, \ldots, s_m\}$$

挑战1：如何全面完整地刻画歌曲

$$Q(u) = < s^*_1, s^*_2, \ldots, s^*_\tau, s^*_{\tau+1}, \ldots, s^*_{\tau+n} > \Rightarrow s^*_{\tau+n+1}?$$

$$s^*_i \in S(1 \le i \le \tau+n)$$

$$|t(\tau+1) - t(\tau)| > \varepsilon$$

$$|t(\tau+i) - t(\tau+i-1)| \le \varepsilon (1 < i \le n)$$

$$R(u) = \{s'_1, s'_2, \ldots, s'_N\}$$

$$s'_i \in S(1 \le i \le N)$$

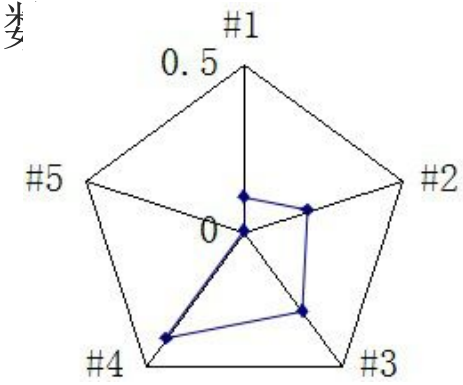
挑战2：如何分析用户收听序列来预测用户行为

背景介绍

- **相关工作：刻画歌曲(挑战1)**
  - Audio Features[Li+@SIGIR'04]
  - Topic Models[Fields+@WOMRAD'10]
    - ✓ 利用用户为歌曲所打的标签构建歌曲对应的文本文档
    - ✓ 使用LDA[Blei+@JMLR'03]等主题模型分析歌曲对应文档，将歌曲表示成若干隐含主题构成的概率分布
    - ✓ 向量的每一维代表一个隐含主题，而每一维的取值表示歌曲隶属于该隐含主题的概率，K表示主题总数

$$\mathbf{s} = (\omega(\mathrm{i},1),\ \omega(\mathrm{i},2),\ldots,\ \omega(\mathrm{i},\mathrm{K}))$$

$$\mathbf{s}_1 = (0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.0)$$



背景介绍

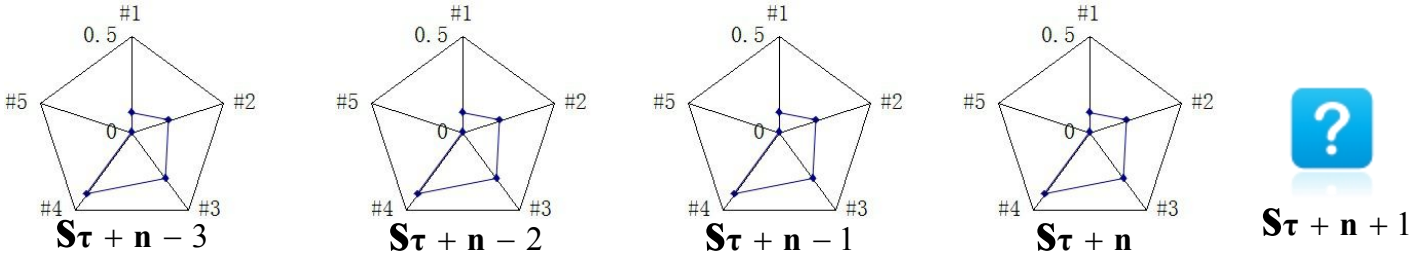
- **相关工作：分析用户行为(挑战2)**
  - 基于用户**即时行为**的音乐推荐 $Q(u)[\tau+n]$
    - ✓Audio-Based[Cano+@MULTIMEDIA'05]
    - ✓Local[Hyung+@IMMM'12]
  - 基于用户**长期行为**的音乐推荐 $Q(u)[1:\tau+n]$
    - ✓UserKNN[Resnick+@CSCW'94]
    - ✓Global[Chordia+@ISMIR'08]
  - 基于用户**中期行为**的音乐推荐 $Q(u)[\tau+1:\tau+n]$
    - ✓PatternMining[Hariri+@RecSys'12]
    - ✓Markov Model[McFee+@ISMIR'11]

背景介绍

- **问题分析**
  - 基于用户中期行为分析的工作得到越来越多的关注，但目前的方法存在如下问题：
    - ✓定性分析
    - ✓考察个别主题的贡献和作用
    - ✓存在对其他用户的依赖
  - 目前未见综合考虑用户三期行为的工作
- **解决思路**
  - 给出一种基于多维时间序列分析的音乐推荐方法，从定量的角度全面地分析用户的行为序列，同时减少对其他用户的依赖
  - 给出一种综合的音乐推荐方法，综合考虑用户三期行为对其未来行为的贡献和作用

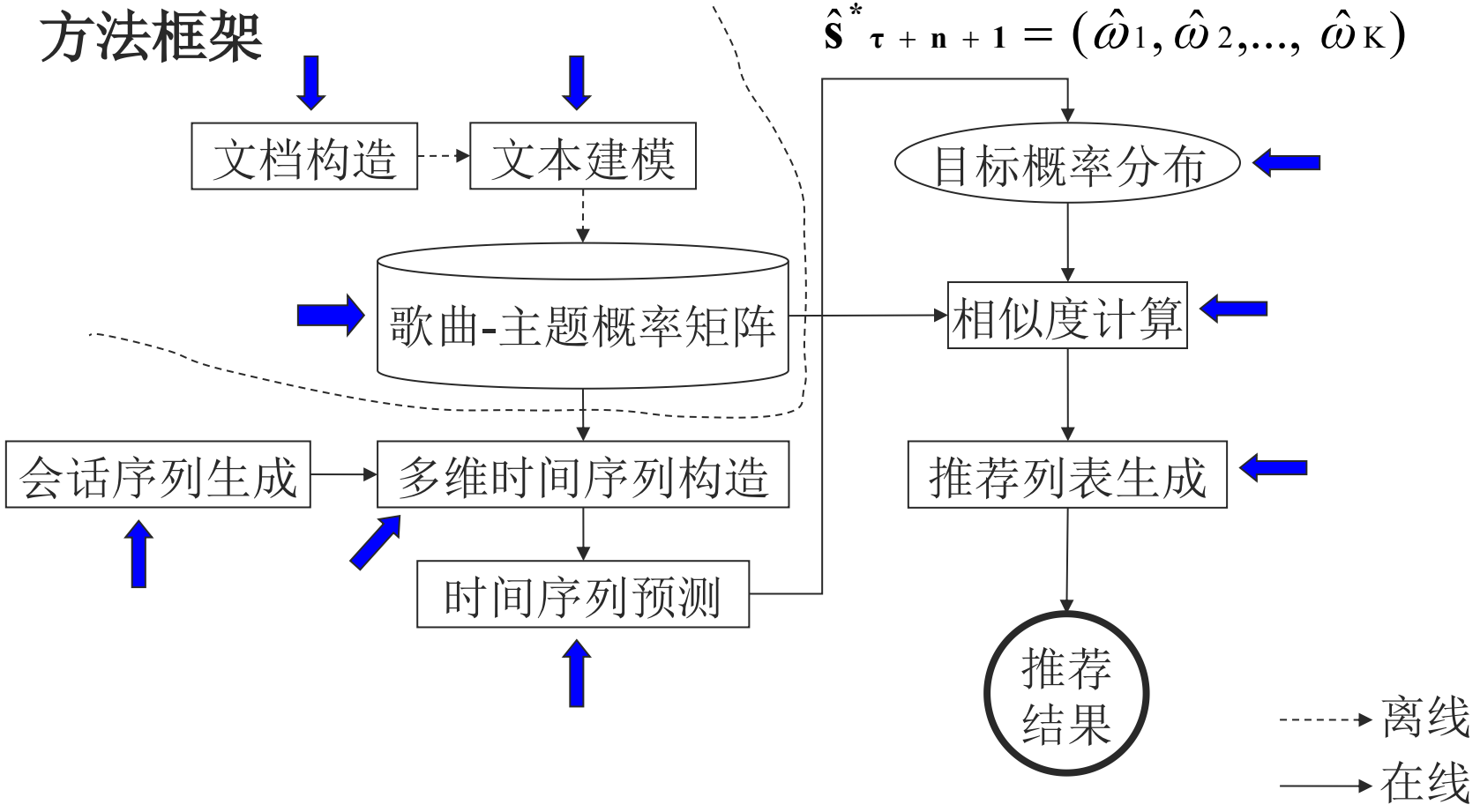
基于多维时序分析的音乐推荐

- **多维时间序列**[Hamilton@Princeton'94]
  - **单变量时间序列**
    - ✓将变量在不同时间上的各个数值按时间先后顺序排列而形成的序列
    - ✓时间序列法是一种定量预测方法
    - ✓x: $<\ldots,x_0,x_1,x_2,x_3,x_4,\ldots>$
  - **多维时间序列**
    - ✓如果任意时刻变量是一个n维向量，比如$\mathbf{X_t}=(X_{1t},X_{2t},\ldots,X_{nt})'$
    - ✓**X:$<\ldots,X_0,X_1,X_2,X_3,X_4\ldots>$**



基于多维时序分析的音乐推荐

- 方法框架



基于多维时序分析的音乐推荐

- **核心思想**
  - 使用**时间序列**分析的方法**定量地**分析并预测用户行为对应的时间序列
  - 分析用户行为在**每一个**主题上的变化趋势，全面考察每一个主题的贡献和作用
- **方法执行**
  - 文本建模：LDA[Blei+@JMLR'03]
  - 时序分析：ARIMA[Box+@Holden-Day'70]
  - 相似度计算: Hellinger距离[Nikulin+@Springer'01]

$$dis(s_i, s_j) = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{K} (\sqrt{s_i(k)} - \sqrt{s_j(k)})^2} \qquad sim(s_i, s_j) = \frac{1}{1 + dis(s_i, s_j)}$$

基于多维时序分析的音乐推荐

- **实验：数据集信息**

用户总数	1530
歌曲总数	24992
歌手总数	5479
最小长度	10
最大长度	30

实验：评测指标

- 预测命中率

$$hitRatio = \frac{\sum_{u \in U} hit(u)}{|U|} \qquad hit(u) = \begin{cases} 1 & , s^*_{\tau+n+1} \in R(u) \\ 0 & , s^*_{\tau+n+1} \notin R(u) \end{cases}$$

用户没收听并不代表用户不喜欢，命中率要求过于严格

- 预测误差

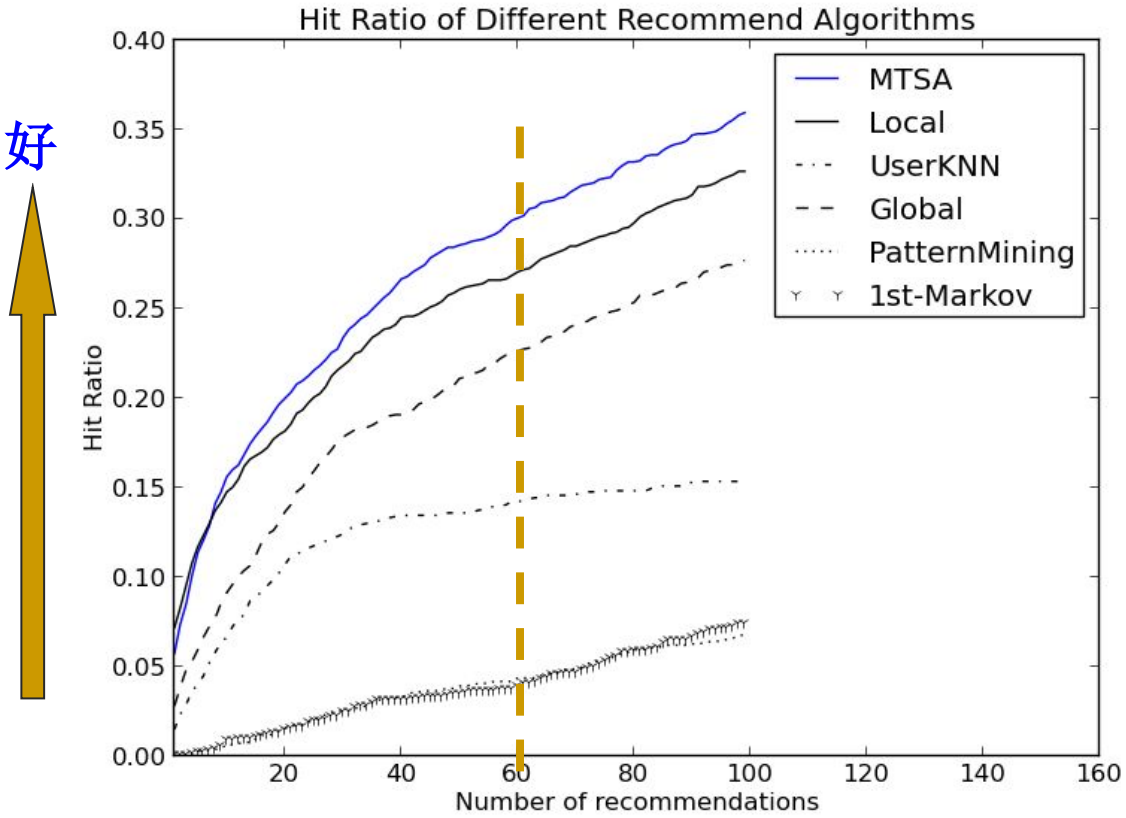
$$RMSE = \sqrt{\frac{\sum_{u \in U} e(u)^2}{|U|}} \qquad e(u) = \frac{\sum_{s \in R(u)} dis(\mathbf{s}, \mathbf{s}^*_{\tau+n+1})}{|R(u)|}$$

命中率高，误差低谓之“好”

（时空效率等指标暂不考虑）

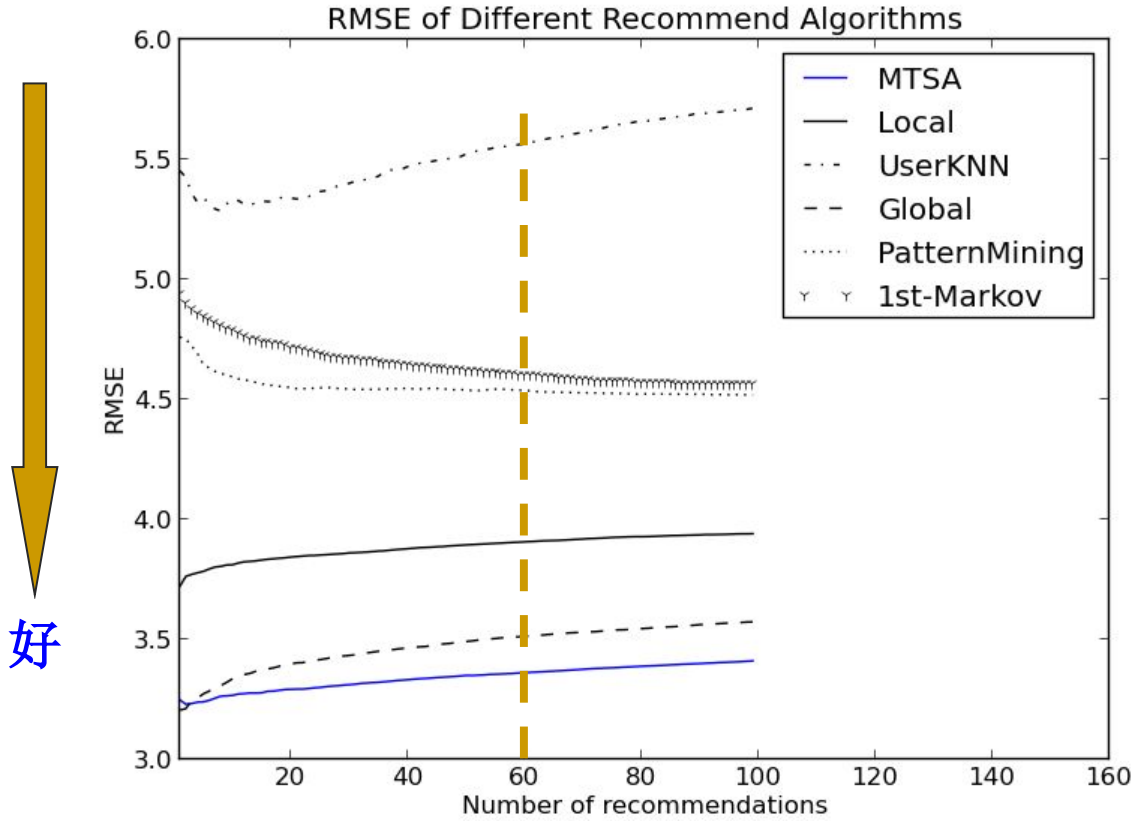
基于多维时序分析的音乐推荐

- **实验结果1：**多维时序方法 VS 参照方法(命中率)



基于多维时序分析的音乐推荐

- **实验结果2：**多维时序方法 VS 参照方法(均方根误差)

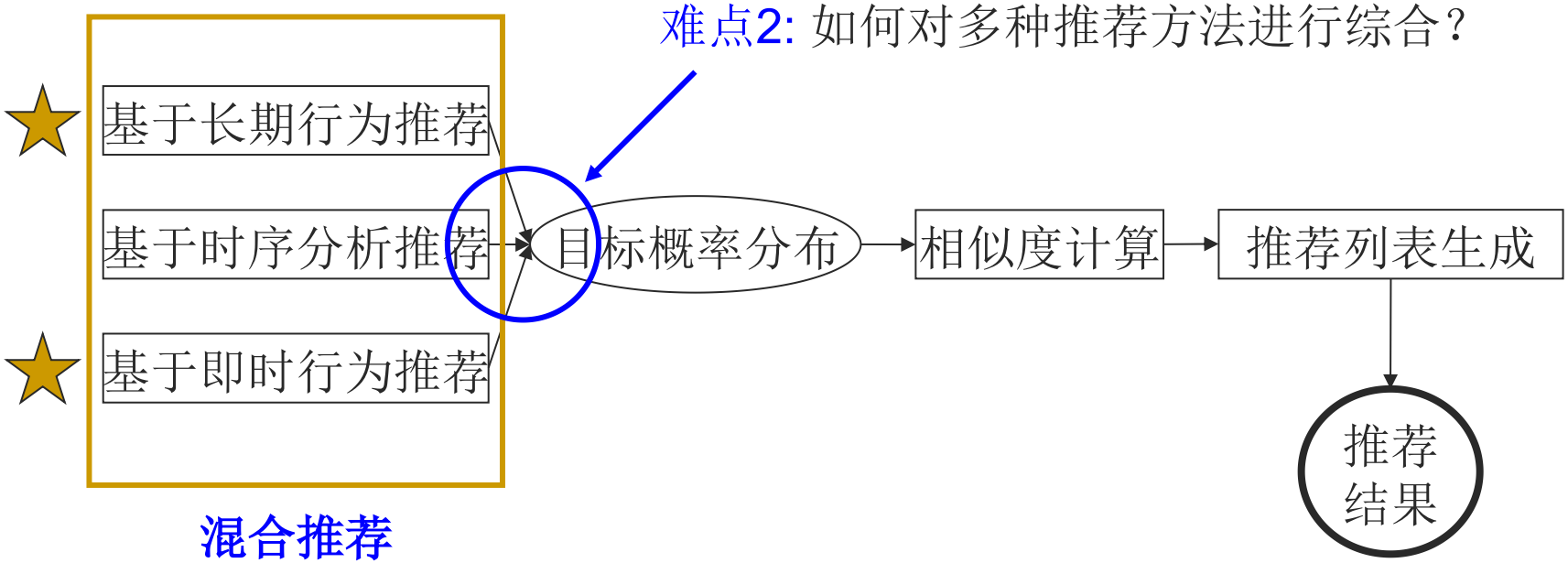


基于用户三期行为的综合音乐推荐

- **方法框架**

难点1: 如何选择待综合的推荐方法？

难点2: 如何对多种推荐方法进行综合？



基于用户三期行为的综合音乐推荐

- **长期性**

  用户长期行为能够在一定程度上反映用户兴趣的倾向性

$$\hat{\omega}_i = \frac{\sum_{j=1}^{j=\tau+n} s^*_j(i)}{\tau + n} (1 \le i \le K)$$

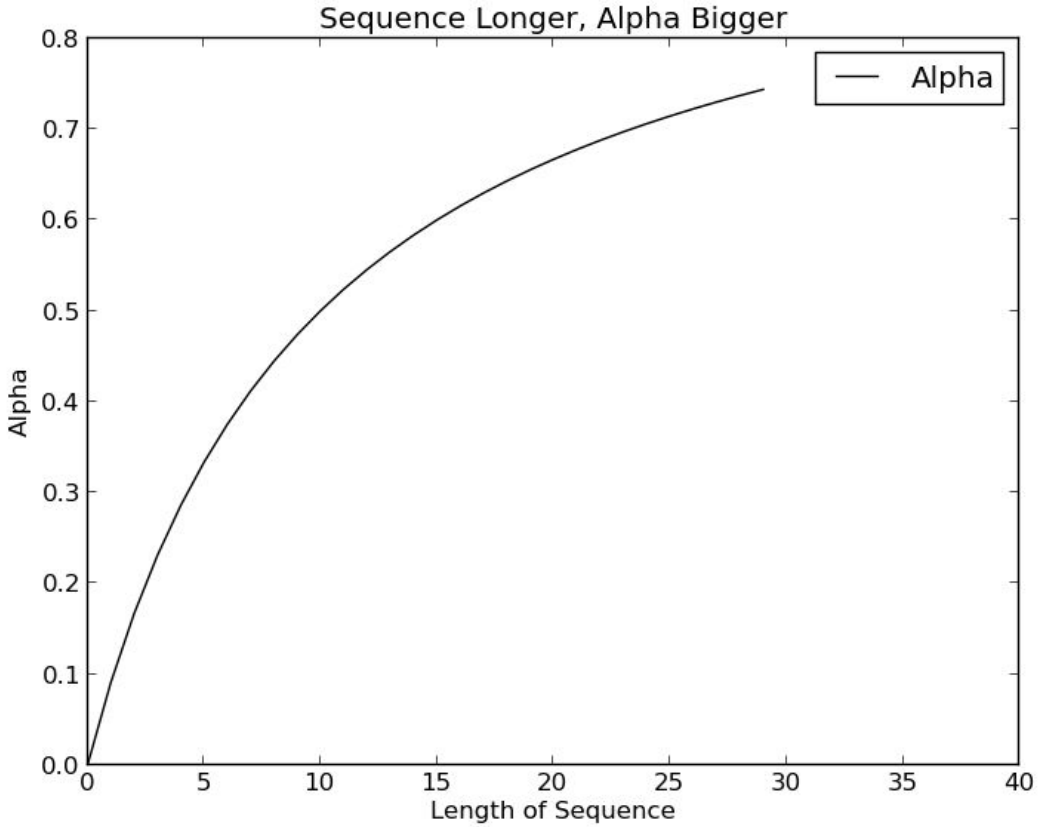
- **即时性**

  用户的状态是稳定的，在短期内不会发生明显地变化

$$\hat{\omega}_i = s^*_{\tau+n}(i) (1 \le i \le K)$$

基于用户三期行为的综合音乐推荐

- **综合思想**
  - 用户的长期行为、中响
  - 用户的中期行为更能
  - 中期行为的权重随着



- **综合策略**

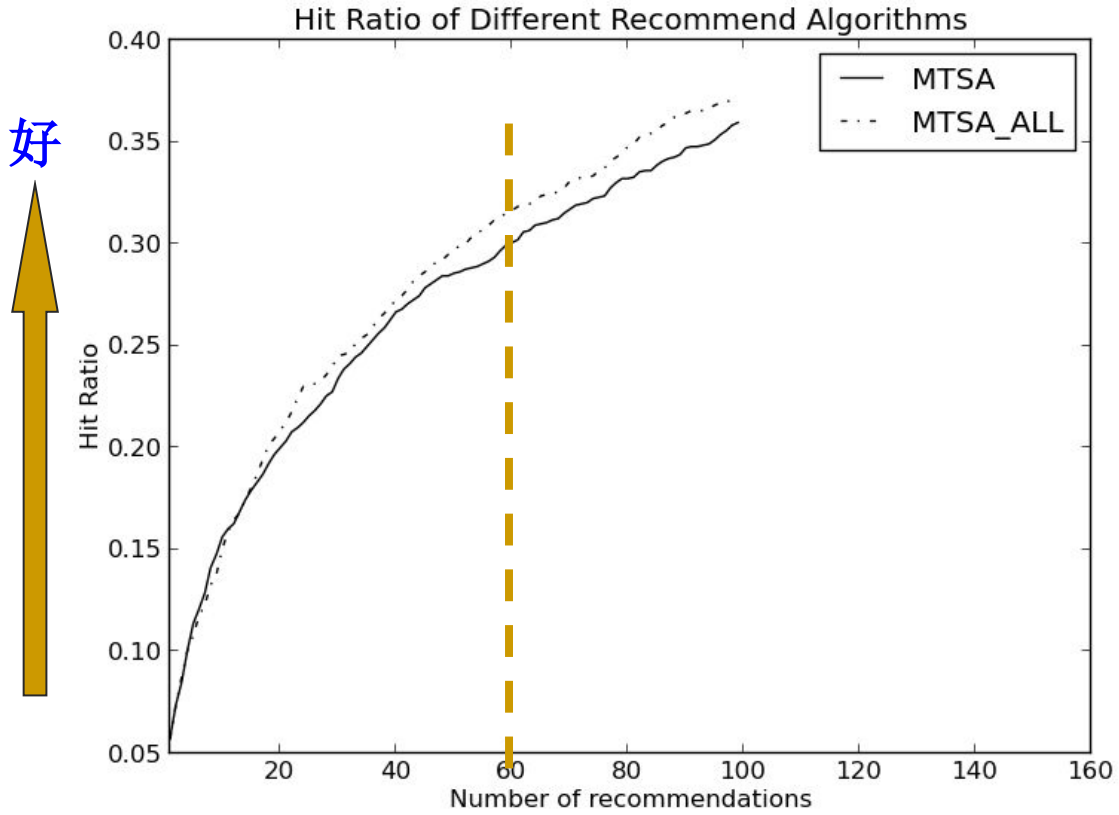
$$\hat{\omega}_i = \alpha\hat{\omega}_{i\_session} + (1-$$

$$\alpha = \frac{len}{len+10}$$

$$\beta = 0.5$$

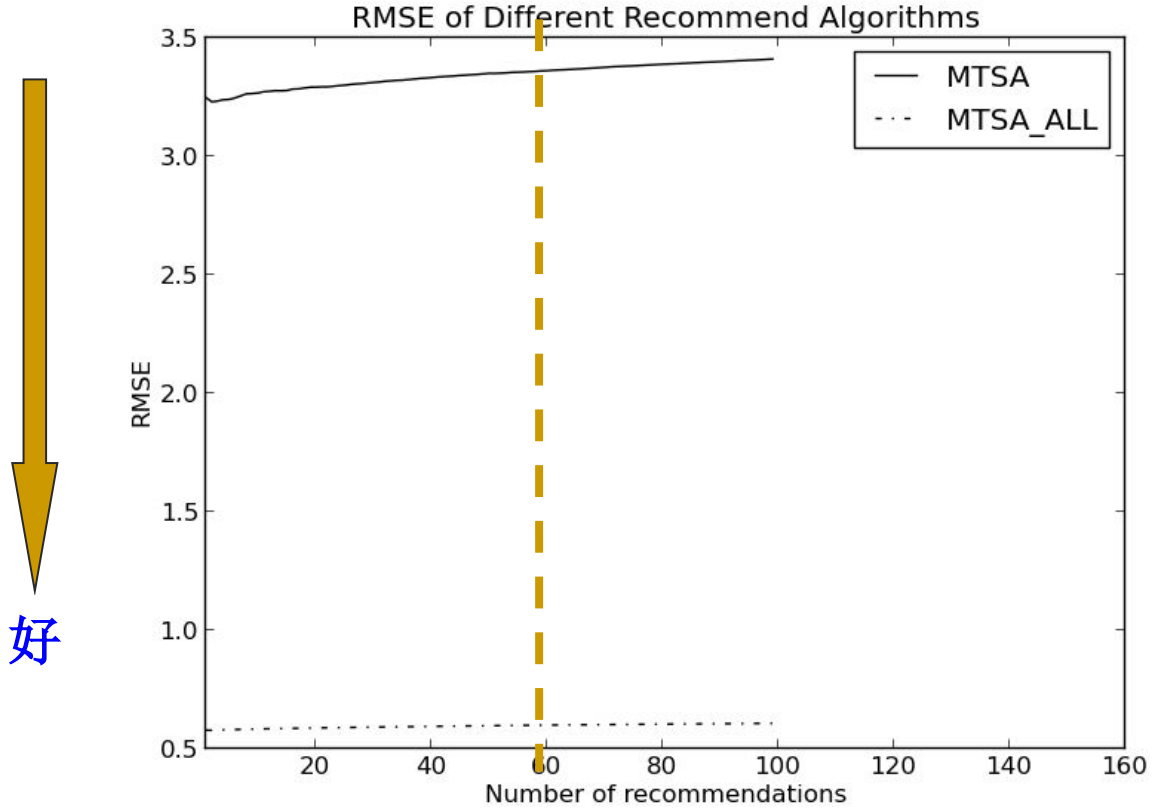
基于用户三期行为的综合音乐推荐

- **实验结果3：**综合方法 VS 基本方法(命中率)



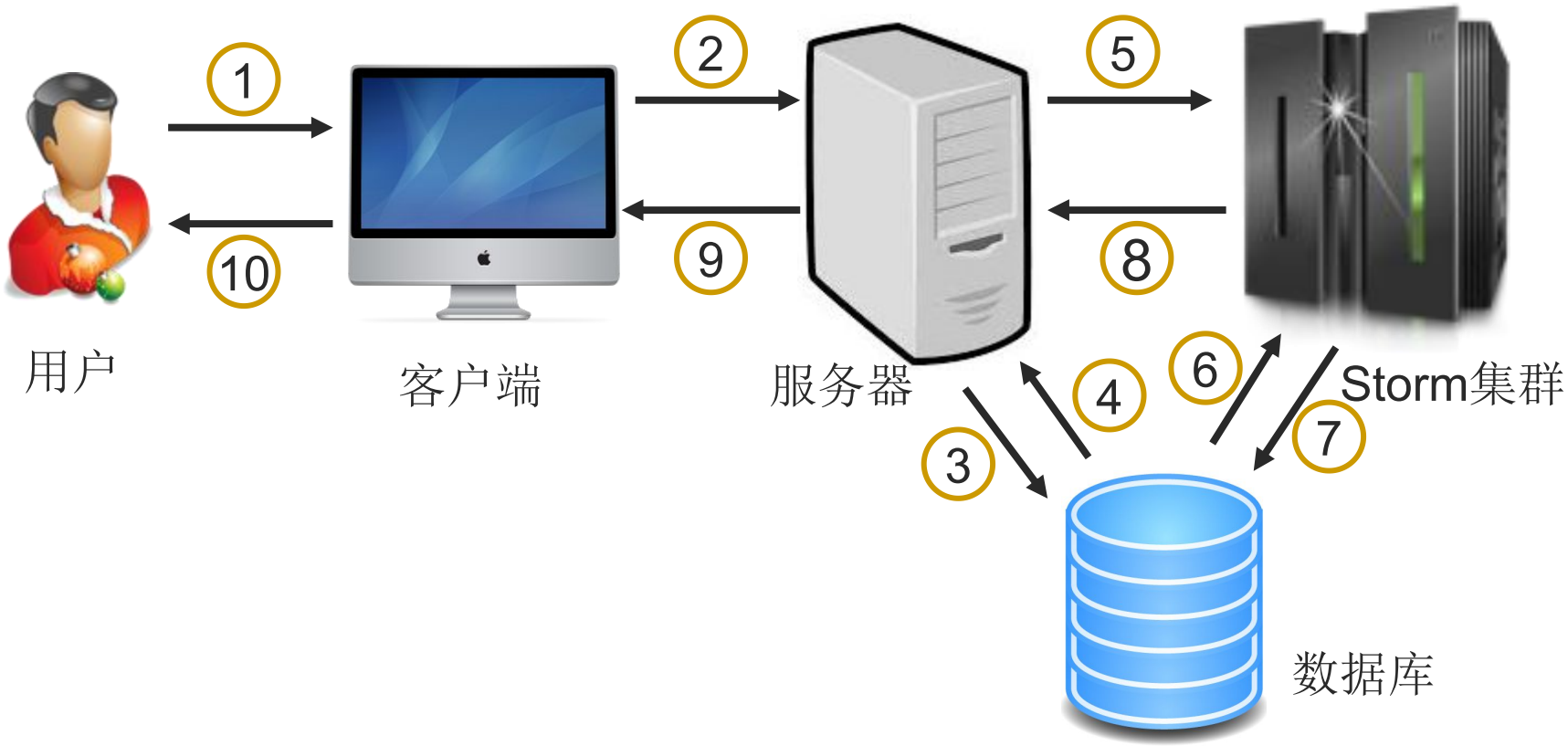
基于用户三期行为的综合音乐推荐

- **实验结果4：**综合方法 VS 基本方法(均方根误差)

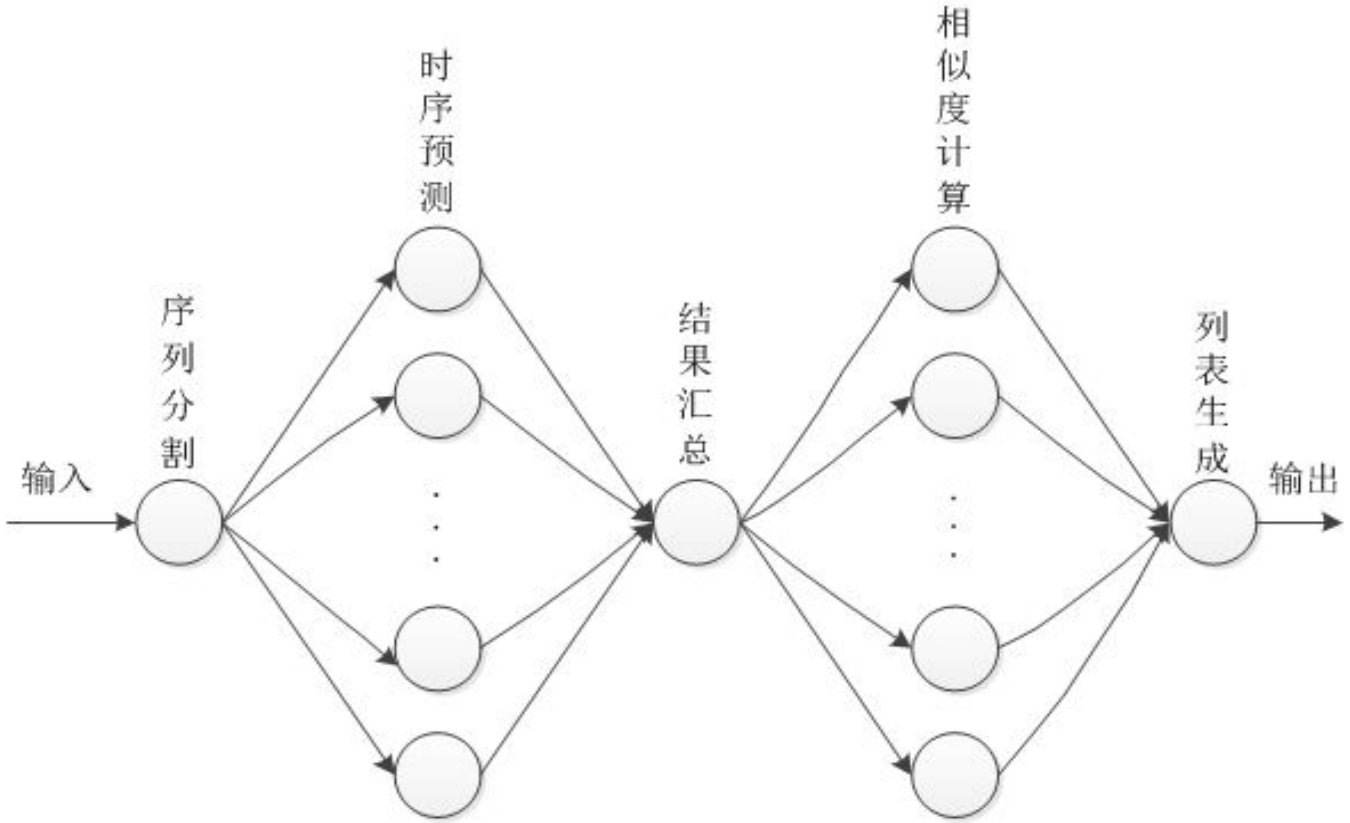


系统实现

- **系统架构**



- 推荐引擎



系统实现

- 系统效果


新用户注册
用户名
密码
确认密码
注册并登录


您好 1 退出
播放：Here's To Never Growing Up-
Here's To Never Growing Up-Avril Lavigne
发行日期：2013-04-09
评分：7.3
出版商：Epic


1
登出
当前歌曲：Here's To Never Growing Up-Avril Lavigne
pop
locking
OK
清除
我要打TAG
继续收听

总结与展望

总结

- 给出了一种基于多维时序分析的个性化音乐推荐方法
- 给出了一种综合推荐方法，充分考虑了用户的长期行为、中期行为和即时行为
- 从不同的角度对所提方法进行评测，验证了所提方法的有效性
- 实现了一个原型系统，验证了所提方法的可行性
- 为所提方法申请专利(已受理)
  - 吕建,徐锋,王守涛. 一种基于多维时间序列分析的个性化音乐推荐系统及其实现方法:中国. 201410077177.1[P]. 2014-03-04.

数据集： http://lastfmseq.sinaapp.com/

源代码： https://github.com/wwssttt/GitRepo/tree/master/Python/experiment

总结与展望

- **展望**
  - 考虑其他本文信息的作用
  - 考虑用户消极行为的作用
  - 考虑各隐含主题对应的时间序列之间的相关性(目前认为它们相互独立)
  - 其他改进

参考文献

[1]Song Y, Dixon S, Pearce M. A survey of music recommendation systems and future perspectives. In 9th international symposium on computer music modelling and retrieval. 2012: 19-22.

[2]Celma O, Lamere P. Music recommendation and discovery revisited. In Proceedings of the fifth ACM conference on Recommender systems. ACM, 2011: 7-8.

[3]Park S E, Lee S, Lee S. Session-based Collaborative Filtering for Predicting the Next Song. In Computers, Networks, Systems and Industrial Engineering (CNSI), 2011 First ACIS/JNU International Conference on. IEEE, 2011: 353-358.

[4]Cano P, Koppenberger M, Wack N. Content-based music audio recommendation. In Proceedings of the 13th annual ACM international conference on Multimedia. ACM, 2005: 211-212.

[5]Hyung Z, Lee M A, Lee K. Music recommendation based on text mining. In the 2nd international conference on advances in information mining and management. 2012: 129-134.

[6]Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994: 175-186.

[7]Chordia P, Godfrey M, Rae A. Extending Content-Based Recommendation: The Case of Indian Classical Music. In the 9th international society for music information retrieval(ISMIR 2008). 2008: 571-576.

[8]Hariri N, Mobasher B, Burke R. Context-aware music recommendation based on latenttopic sequential patterns. In Proceedings of the sixth ACM conference on Recommender systems. ACM, 2012: 131-138.

[9]McFee B, Lanckriet G R G. The Natural Language of Playlists. In the 12th international society for music information retrieval(ISMIR 2011). 2011: 537-542.

[10]Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. In the Journal of machine Learning research, 2003, 3: 993-1022.

[11]Box, George and Jenkins, Gwilym (1970) Time series analysis: Forecasting and control, San Francisco: Holden-Day, 1970.

[12]Li Q, Kim B M, Guan D H. A music recommender based on audio features[C]//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004: 532-533.

[13]Fields B, Rhodes C, d'Inverno M. Using song social tags and topic models to describe and compare playlists[C]//1st Workshop On Music Recommendation And Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain. 2010.

[14]James D. Hamilton. Time Series Analysis, Princeton, New Jersey.

[15]Nikulin, M.S. (2001), "Hellinger distance", in Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4

谢谢！！

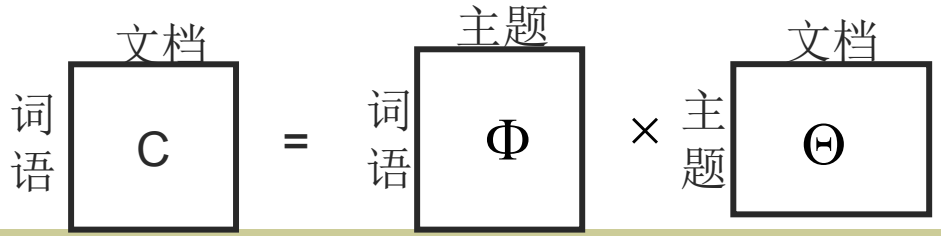
附录

- LDA
- 多维时间序列
- ARIMA
- Hellinger距离
- 不同文本模型之间的比较
- 最大可分析序列长度
- 离散概率分布

附录

- **LDA**
  - **描述：**是一种主题模型，它可以将文档集中每篇文档的主题按照概率分布的形式给出
  - **假设：**文档由若干主题构成，主题由若干词汇表征，通过主题将文档和词汇联系起来
  - **输入：**文档集合、主题数目K
  - **输出：**文档-主题矩阵(文档属于各个主题的概率)
    主题-词汇矩阵(词汇表征各个主题的概率)

$$p(\text{词语}\,|\,\text{文档}) = \sum_{\text{主题}} p(\text{词语}\,|\,\text{主题}) \times p(\text{主题}\,|\,\text{文档})$$

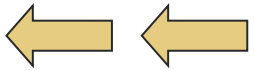


附录

多维时间序列

- 设$T=\{...-2,-1,0,1,2...\}$是一个足标集，对任意固定的$t\in T$, Y_t是随机变量，$t\in T$的全体$\{Y_t:t\in T\}$称为T上的单变量时间序列，记为$\{Y_t\}$
- 如果 Y_t是n维随机变量，即 $Y_t=(Y_{1t},Y_{2t},...,Y_{nt})'$ 那么 $t\in T$的全体$\{Y_t:t\in T\}$称为T上的n维时间序列，记为 $\{Y_t\}$

- 要理解多维时间序列需要注意两点：
  - ✓ 对于任意固定时刻，Y_t是n维随机变量
  - ✓ 对每个分量来说，$\{Y_{it}\}$是一个单变量时间序列



附录

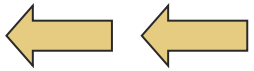
- **ARIMA**
  - 差分整合移动平均自回归模型是时间序列分析和预测的一种非常常用的方法，由Box、Jenkins等人于70年代提出，因此又称为Box-Jenkins模型
  - 该模型包括自回归模型和滑动平均模型两部分，自回归模型描述的是当前值与历史值之间的关系，滑动平均模型描述的是自回归部分的误差累计
  - 该模型认为时间序列在时刻t的取值y_t与其其前面p个变量的取值以及q个随机误差项的取值相关

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

- **Hellinger距离**
  - 对于两个离散概率分布$P=(p_1,p_2,...,p_n)$和$Q=(q_1,q_2,...,q_n)$,其Hellinger距离可由下式计算得来
  - 两个概率分布距离越小越相似，反之越不相似

$$H(P,Q)=\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{p_i}-\sqrt{q_i})^2}$$

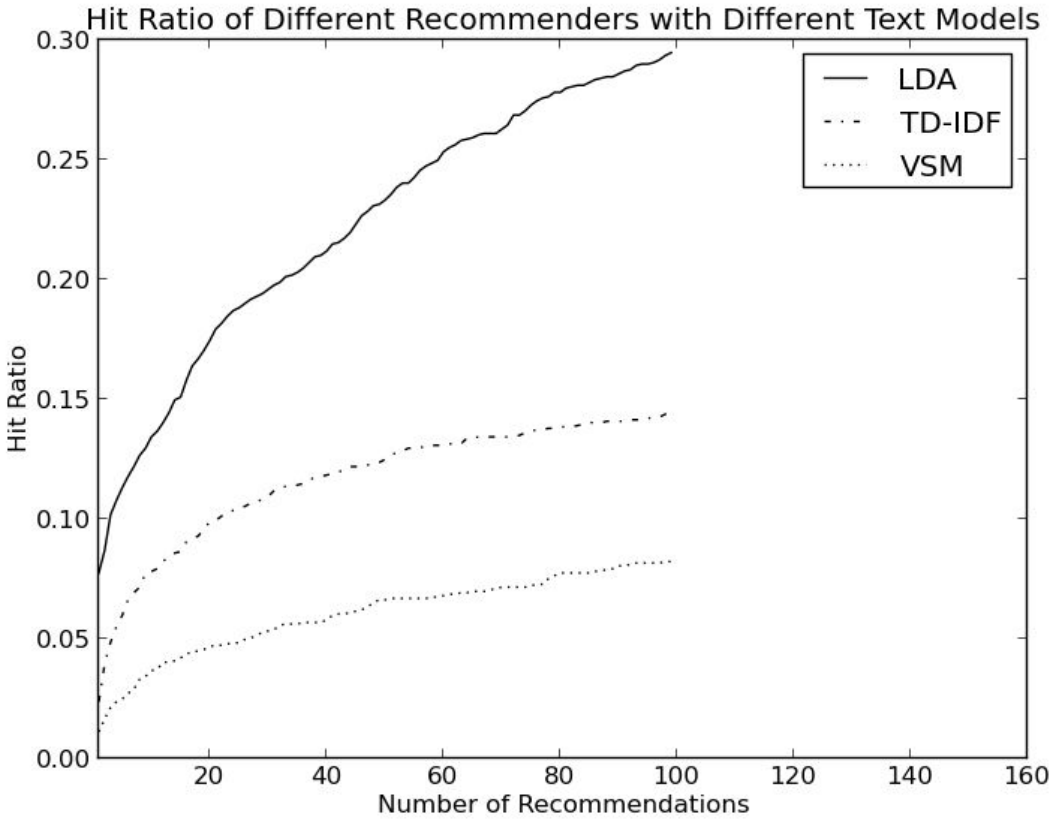
$$sim(s_i,s_j)=\frac{1}{1+H(s_i,s_j)}$$



附录

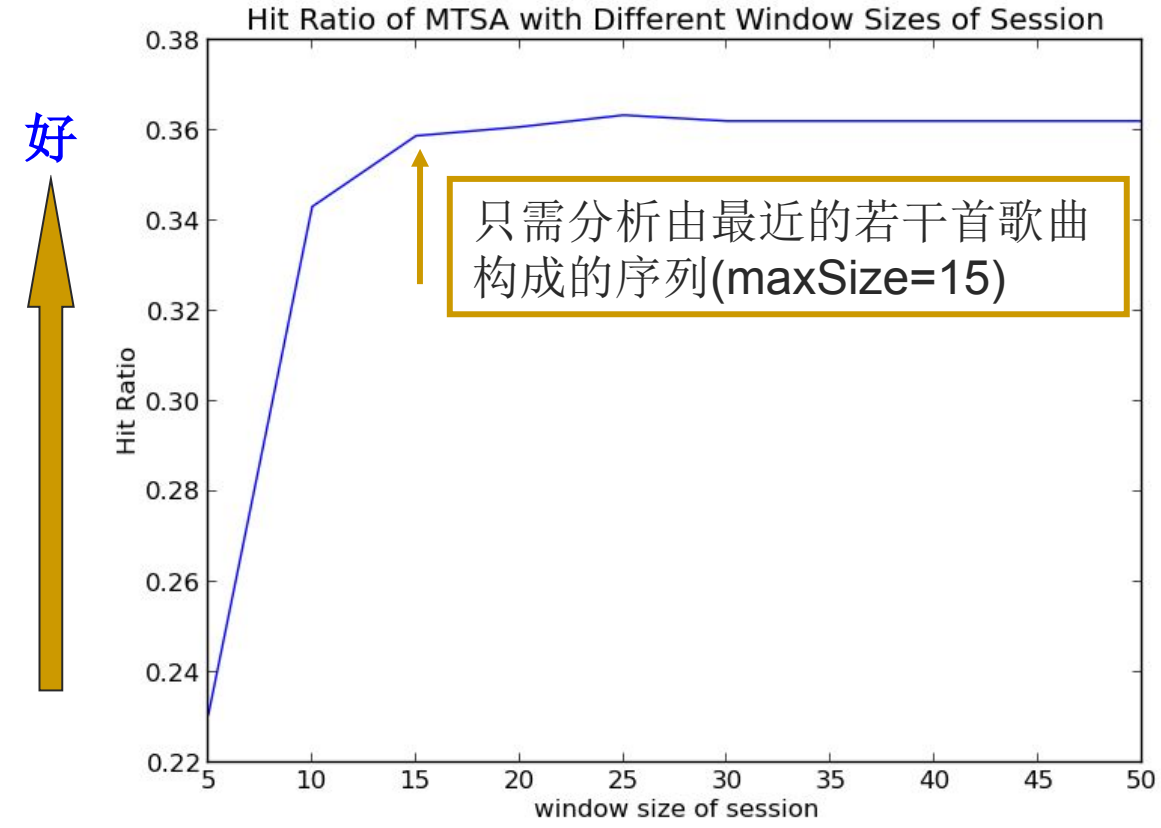
- 不同文本建模方法比较

好



附录

- 最大可分析序列长度



- **离散概率分布**

1. 离散型随机变量的概念

定义 若随机变量 X 的可能取值是有限多个或无穷可列多个，则称 X 为离散型随机变量

描述离散型随机变量的概率特性常用它的概率分布或分布律，即

$$P(X = x_k) = p_k,\ k = 1,2,\cdots$$

概率分布的性质：

- $p_k \geq 0,\ k = 1,2,\cdots$
- $\sum_{k=1}^{\infty} p_k = 1$