A Latent Semantic Pattern Recognition Strategy for an Untrivial Targeted Advertising

Roberto Saia Ludovico Boratto Salvatore Carta

Department of Mathematics and Computer Science
University of Cagliari, Italy

IEEE 2015 - 4th International Congress on Big Data
June 27 - July 02, 2015 New York, USA

Outline

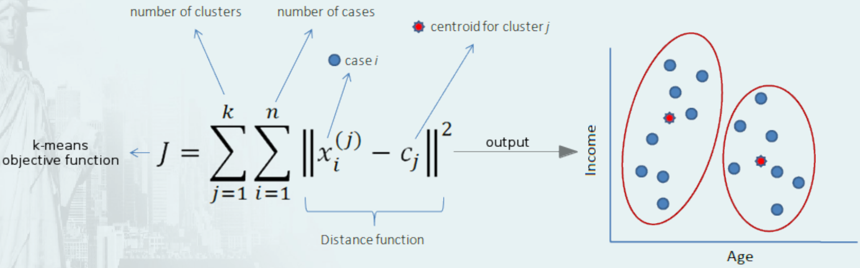
INTRODUCTION

BEHAVIORAL TARGETING

- ✓ *Behavioral targeting* is the strategy used to identify sets of users who share common properties
- ✓ Classic partitioning algorithms (e.g.,*k-means*) cannot take in account the semantic of the user behavior (*Gong et al.* [2])

BEHAVIORAL TARGETING

- ✓ *Behavioral targeting* is the strategy used to identify sets of users who share common properties
- ✓ Classic partitioning algorithms (e.g.,*k-means*) cannot take in account the semantic of the user behavior (*Gong et al.* [2])



RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- In the literature strategies that take into account the semantic aspect have been proposed:
  - *Tu and Lu* [5], based on query semantic analysis
  - *Gong et al.* [2], based on queries and clicks analysis
- However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

RELATED WORK AND OPEN PROBLEMS

- ✓ In the literature strategies that take into account the semantic aspect have been proposed:
  - ✓ *Tu and Lu* [5], based on query semantic analysis
  - ✓ *Gong et al.* [2], based on queries and clicks analysis
- ✓ However, some problems remain open:
  1. The **reliability** of a semantic query analysis (*Spink et al.* [4])
     - ✓ Nearly 50% of the times users reformulate their queries
  2. The **interpretability** of the segments (*Nairn and Bottomley* [3])
     - ✓ A segmentation is either trivial or too complex to analyze (too many features involved)
  3. The **preference stability** (*Burke and Ramezani* [1])
     - ✓ Users tend to choose items of the same classes (e.g., movies of the same genres)

PROPOSED SOLUTION

- ✓ Semantic analysis of the description of the items positively evaluated by the users
  - ✓ This solve the **reliability** of the semantic query analysis problem (*step 1*)
- ✓ Definition of novel binary filters that describe which terms (Wordnet[1] synsets) characterize each class of items
  - ✓ This solves the segments **interpretability** problem (*steps 2 and 3*)
- ✓ Matching between the user profiles and the binary filters, in order to detect non trivial user segments
  - ✓ This solves the problem related to the **preference stability** (*steps 4 and 5*)

[1] A lexical database that groups words into set of synonyms called synsets

PROPOSED SOLUTION

- ✓ Semantic analysis of the description of the items positively evaluated by the users
  - ✓ This solve the **reliability** of the semantic query analysis problem (*step 1*)
- ✓ Definition of novel binary filters that describe which terms (Wordnet[1] synsets) characterize each class of items
  - ✓ This solves the segments **interpretability** problem (*steps 2 and 3*)
- ✓ Matching between the user profiles and the binary filters, in order to detect non trivial user segments
  - ✓ This solves the problem related to the **preference stability** (*steps 4 and 5*)

1
Items Description
Preprocessing

[1] A lexical database that groups words into set of synonyms called synsets

PROPOSED SOLUTION

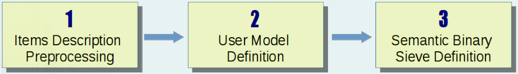
- ✓ Semantic analysis of the description of the items positively evaluated by the users
  - ✓ This solve the **reliability** of the semantic query analysis problem (*step 1*)
- ✓ Definition of novel binary filters that describe which terms (Wordnet[1] synsets) characterize each class of items
  - ✓ This solves the segments **interpretability** problem (*steps 2 and 3*)
- ✓ Matching between the user profiles and the binary filters, in order to detect non trivial user segments
  - ✓ This solves the problem related to the **preference stability** (*steps 4 and 5*)

1
Items Description
Preprocessing

[1] A lexical database that groups words into set of synonyms called synsets

PROPOSED SOLUTION

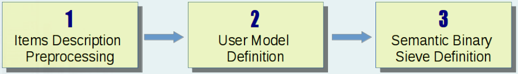
- ✓ Semantic analysis of the description of the items positively evaluated by the users
  - ✓ This solve the **reliability** of the semantic query analysis problem (*step 1*)
- ✓ Definition of novel binary filters that describe which terms (Wordnet[1] synsets) characterize each class of items
  - ✓ This solves the segments **interpretability** problem (*steps 2 and 3*)
- ✓ Matching between the user profiles and the binary filters, in order to detect non trivial user segments
  - ✓ This solves the problem related to the **preference stability** (*steps 4 and 5*)



[1] A lexical database that groups words into set of synonyms called synsets

PROPOSED SOLUTION

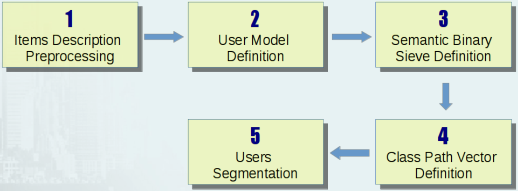
- Semantic analysis of the description of the items positively evaluated by the users
  - This solve the **reliability** of the semantic query analysis problem (*step 1*)
- Definition of novel binary filters that describe which terms (Wordnet[1] synsets) characterize each class of items
  - This solves the segments **interpretability** problem (*steps 2 and 3*)
- Matching between the user profiles and the binary filters, in order to detect non trivial user segments
  - This solves the problem related to the **preference stability** (*steps 4 and 5*)



[1] A lexical database that groups words into set of synonyms called synsets

PROPOSED SOLUTION

- ✓ Semantic analysis of the description of the items positively evaluated by the users
  - ✓ This solve the **reliability** of the semantic query analysis problem (*step 1*)
- ✓ Definition of novel binary filters that describe which terms (Wordnet[1] synsets) characterize each class of items
  - ✓ This solves the segments **interpretability** problem (*steps 2 and 3*)
- ✓ Matching between the user profiles and the binary filters, in order to detect non trivial user segments
  - ✓ This solves the problem related to the **preference stability** (*steps 4 and 5*)



[1] A lexical database that groups words into set of synonyms called synsets

IMPLEMENTATION

1. ITEMS DESCRIPTION PREPROCESSING

- ✓ We preprocess the textual description of each item, to remove the useless elements (i.e., punctuation marks and stop-words)
- ✓ We determine the lemma of each word (e.g., use *good* instead of *better*, and *walk* instead of *walking*)
- ✓ We assign to them the best sense (i.e., **1**=noun, **2**=verb)

Input

In the near future, a computer hacker named Neo discovers that all life on Earth may be nothing more than an elaborate illusion

Output

future/**1** computer/**1** hacker/**1** nam/**2** neo/**1** discover/**2** life/**1** earth/**1** illusion/**1**

1. ITEMS DESCRIPTION PREPROCESSING

- ✓ We preprocess the textual description of each item, to remove the useless elements (i.e., punctuation marks and stop-words)
- ✓ We determine the lemma of each word (e.g., use *good* instead of *better*, and *walk* instead of *walking*)
- ✓ We assign to them the best sense (i.e., **1**=noun, **2**=verb)

Input

In the near future, a computer hacker named Neo discovers that all life on Earth may be nothing more than an elaborate illusion

Output

future/**1** computer/**1** hacker/**1** nam/**2** neo/**1** discover/**2** life/**1** earth/**1** illusion/**1**

1. ITEMS DESCRIPTION PREPROCESSING

- ✓ We preprocess the textual description of each item, to remove the useless elements (i.e., punctuation marks and stop-words)
- ✓ We determine the lemma of each word (e.g., use *good* instead of *better*, and *walk* instead of *walking*)
- ✓ We assign to them the best sense (i.e., **1**=noun, **2**=verb)

Input

In the near future, a computer hacker named Neo discovers that all life on Earth may be nothing more than an elaborate illusion

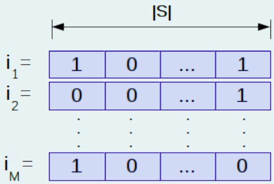
Output

future/**1** computer/**1** hacker/**1** nam/**2** neo/**1** discover/**2** life/**1** earth/**1** illusion/**1**

2. USER MODEL DEFINITION

$$m_u[w] = \begin{cases} 1, & \textit{if } ds_i[w] = 1 \\ m_u[w], & \textit{otherwise} \end{cases}$$

- ✓ For each user $u \in U$, we consider the set of items I_u she/he likes

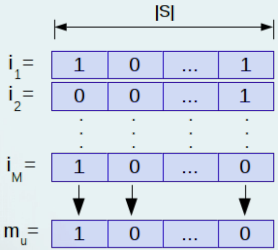


[2] $|S|$ denotes the cardinality of distinct synsets

2. USER MODEL DEFINITION

$$m_u[w] = \begin{cases} 1, & \textit{if } ds_i[w] = 1 \\ m_u[w], & \textit{otherwise} \end{cases}$$

- ✓ For each user $u \in U$, we consider the set of items I_u she/he likes
- ✓ Then build a model m_u that describes which synsets[2] appear in the semantic description ds_i of each item i

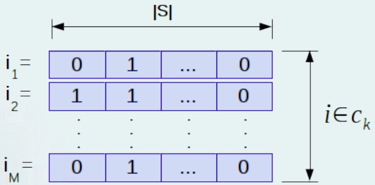


[2] $|S|$ denotes the cardinality of distinct synsets

3. SEMANTIC BINARY SIEVE DEFINITION

$$b_k[w] = \begin{cases} 1, & \textit{if } ds_i[w] = 1 \wedge i \in c_k \\ b_k[w], & \textit{otherwise} \end{cases}$$

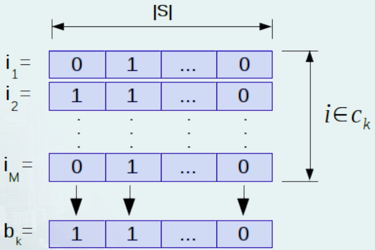
- ✓ For each class $c \in C$, we create a binary vector that reports which synsets are relevant for that class



3. SEMANTIC BINARY SIEVE DEFINITION

$$b_k[w] = \begin{cases} 1, & \text{if } ds_i[w] = 1 \wedge i \in c_k \\ b_k[w], & \text{otherwise} \end{cases}$$

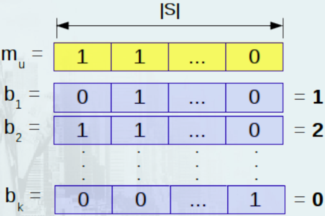
- ✓ For each class $c \in C$, we create a binary vector that reports which synsets are relevant for that class
- ✓ Then we stored each vector b_k in a set $B = \{b_1, \ldots, b_K\}$ (note that $|B| = |C|$)



4. CLASS PATH VECTOR DEFINITION

$$b_k[w] \ominus m_u[w] = \begin{cases} r_u[k]++, & \textit{if } m_u[w] = 1 \wedge b_k[w] = 1 \\ r_u[k], & \textit{otherwise} \end{cases}$$

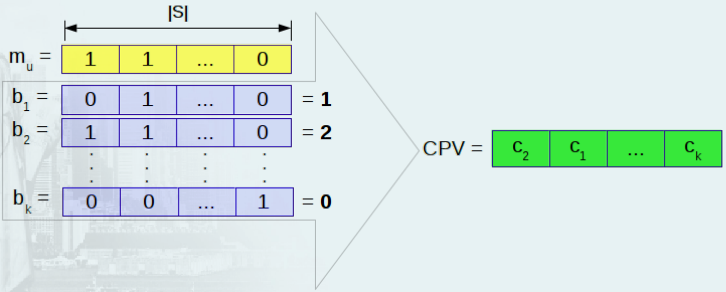
- ✓ We compare ($\ominus$ operator) the *user models* $M = \{m_1, \ldots, m_N\}$ and the *SBS* vectors $B = \{b_1, \ldots, b_K\}$



4. CLASS PATH VECTOR DEFINITION

$$b_k[w] \ominus m_u[w] = \begin{cases} r_u[k]++, & \textit{if } m_u[w] = 1 \wedge b_k[w] = 1 \\ r_u[k], & \textit{otherwise} \end{cases}$$

- ✓ We compare ($\ominus$ operator) the *user models* $M = \{m_1, \ldots, m_N\}$ and the *SBS* vectors $B = \{b_1, \ldots, b_K\}$
- ✓ Then build a *Class Path Vector* (*CPV*) for each user, which contains the class relevance in decreasing order

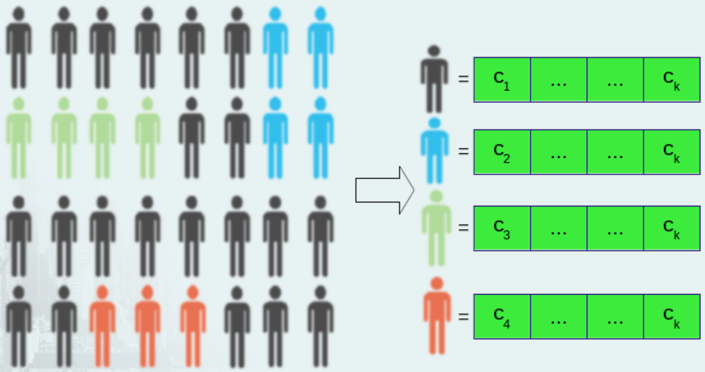


5. USER SEGMENTATION

✓ We segment the users by querying the CPV models previously built

5. USER SEGMENTATION

- ✓ We segment the users by querying the CPV models previously built
- ✓ Then place each user in the class with the highest value of relevance score



EXPERIMENTS

FRAMEWORK

- ✓ The experiments were performed using 2 real-world datasets: the *Yahoo! Webscope Movie (R4)*[3], and the *Movielens 10M*[4]
- ✓ We compare our strategy with:
  - ✓ the k-means approach
  - ✓ the *native classification* of the items, obtained by selecting for each user the most preferred class in the real dataset
- ✓ The aim is to study the *users distribution* (i.e., the ability to group the users in a non-trivial way) and the *segments composition* (i.e., the ability to group the users in a pertinent way)

[3] http://www.cs.umd.edu/~mount/pubs.html
[4] http://grouplens.org/datasets/movielens/

FRAMEWORK

- The experiments were performed using 2 real-world datasets: the *Yahoo! Webscope Movie (R4)*[3], and the *Movielens 10M*[4]
- We compare our strategy with:
  - the k-means approach
  - the *native classification* of the items, obtained by selecting for each user the most preferred class in the real dataset
- The aim is to study the *users distribution* (i.e., the ability to group the users in a non-trivial way) and the *segments composition* (i.e., the ability to group the users in a pertinent way)

[3] http://www.cs.umd.edu/~mount/pubs.html
[4] http://grouplens.org/datasets/movielens/

FRAMEWORK

- The experiments were performed using 2 real-world datasets: the *Yahoo! Webscope Movie (R4)*[3], and the *Movielens 10M*[4]
- We compare our strategy with:
  - the k-means approach
  - the *native classification* of the items, obtained by selecting for each user the most preferred class in the real dataset
- The aim is to study the *users distribution* (i.e., the ability to group the users in a non-trivial way) and the *segments composition* (i.e., the ability to group the users in a pertinent way)

[3] http://www.cs.umd.edu/~mount/pubs.html
[4] http://grouplens.org/datasets/movielens/

FRAMEWORK

- ✓ The experiments were performed using 2 real-world datasets: the *Yahoo! Webscope Movie (R4)*[3], and the *Movielens 10M*[4]
- ✓ We compare our strategy with:
  - ✓ the k-means approach
  - ✓ the *native classification* of the items, obtained by selecting for each user the most preferred class in the real dataset
- ✓ The aim is to study the *users distribution* (i.e., the ability to group the users in a non-trivial way) and the *segments composition* (i.e., the ability to group the users in a pertinent way)

---

[3] http://www.cs.umd.edu/~mount/pubs.html
[4] http://grouplens.org/datasets/movielens/

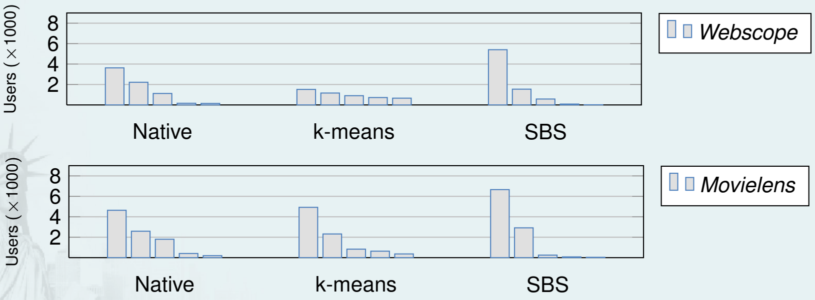
FRAMEWORK

- ✓ The experiments were performed using 2 real-world datasets: the *Yahoo! Webscope Movie (R4)*[3], and the *Movielens 10M*[4]
- ✓ We compare our strategy with:
  - ✓ the k-means approach
  - ✓ the *native classification* of the items, obtained by selecting for each user the most preferred class in the real dataset
- ✓ The aim is to study the *users distribution* (i.e., the ability to group the users in a non-trivial way) and the *segments composition* (i.e., the ability to group the users in a pertinent way)

[3] http://www.cs.umd.edu/~mount/pubs.html

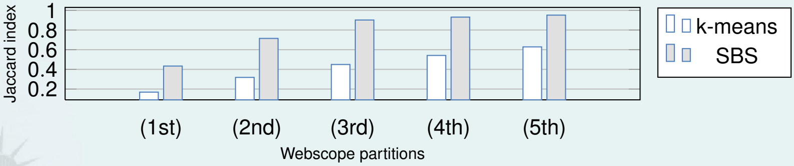
[4] http://grouplens.org/datasets/movielens/

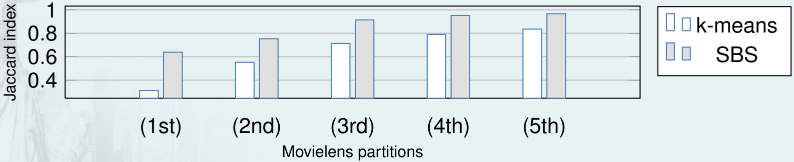
USERS DISTRIBUTION



The results show that *k-means* segments the users in a quite uniform way (*Webscope*), or like the native classification (*Movielens*), since it is not able to discover any implicit preferences, differently from our approach

SEGMENTS COMPOSITION





The relevance of our partitioning is proved by the strong overlap between with the users in the native classification, differently from those classified by *k-means*

CONCLUSIONS

CONCLUSIONS

- In this paper we presented an approach for the user segmentation, solving three open problems:
  1. **Source reliability**: solved by performing a semantic analysis of the description of the items positively evaluated by the users
  2. **Interpretability**: solved by defining and comparing two sets of binary filters (i.e., *User Models* and *SBS*)
  3. **Preference stability**: solved by generating (through the *CPV*) a non-trivial but relevant user segmentation

CONCLUSIONS

- In this paper we presented an approach for the user segmentation, solving three open problems:
  1. **Source reliability**: solved by performing a semantic analysis of the description of the items positively evaluated by the users
  2. **Interpretability**: solved by defining and comparing two sets of binary filters (i.e., *User Models* and *SBS*)
  3. **Preference stability**: solved by generating (through the *CPV*) a non-trivial but relevant user segmentation

CONCLUSIONS

- In this paper we presented an approach for the user segmentation, solving three open problems:
  1. **Source reliability**: solved by performing a semantic analysis of the description of the items positively evaluated by the users
  2. **Interpretability**: solved by defining and comparing two sets of binary filters (i.e., *User Models* and *SBS*)
  3. **Preference stability**: solved by generating (through the *CPV*) a non-trivial but relevant user segmentation

CONCLUSIONS

- In this paper we presented an approach for the user segmentation, solving three open problems:
  1. **Source reliability**: solved by performing a semantic analysis of the description of the items positively evaluated by the users
  2. **Interpretability**: solved by defining and comparing two sets of binary filters (i.e., *User Models* and *SBS*)
  3. **Preference stability**: solved by generating (through the *CPV*) a non-trivial but relevant user segmentation

FUTURE WORK

- ✓ Future work will test the ability of the proposed approach to characterize clusters of users who bought items semantically related
- ✓ This approach allows us to perform group recommendations, by suggesting items to groups of “semantically related” users
- ✓ Considering that some users could be assigned to a class in which they have not evaluated any items, this can generate serendipitous recommendations

FUTURE WORK

- ✓ Future work will test the ability of the proposed approach to characterize clusters of users who bought items semantically related
- ✓ This approach allows us to perform group recommendations, by suggesting items to groups of “semantically related” users
- ✓ Considering that some users could be assigned to a class in which they have not evaluated any items, this can generate serendipitous recommendations

FUTURE WORK

- ✓ Future work will test the ability of the proposed approach to characterize clusters of users who bought items semantically related
- ✓ This approach allows us to perform group recommendations, by suggesting items to groups of “semantically related” users
- ✓ Considering that some users could be assigned to a class in which they have not evaluated any items, this can generate serendipitous recommendations

REFERENCES

1. **R. D. Burke and M. Ramezani.**
Matching recommendation technologies and domains.
In Recommender Systems Handbook, pages 367-386. Springer, 2011.

2. **X. Gong, X. Guo, R. Zhang, X. He, and A. Zhou.**
Search behavior based latent semantic user segmentation for advertising targeting.
In Data Mining (ICDM), 2013 IEEE 13th International Conference on, pages 211-220, Dec 2013.

3. **A. Nairn and P. Bottomley.**
Something approaching science? cluster analysis procedures in the crm era.
In Proceedings of the 2002 Academy of Marketing Science (AMS) Annual Conference, Developments in Marketing Science: Proceedings of the Academy of Marketing Science, pages 120-120. Springer International Publishing, 2003.

4. **A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic.**
From e-sex to e-commerce: Web search changes.
Computer, 35(3):107-109, Mar. 2002.

5. **S. Tu and C. Lu.**
Topic-based user segmentation for online advertising with latent dirichlet allocation.
In Proceedings of the 6th International Conference on Advanced Data Mining and Applications - Volume Part II, ADMA'10, pages 259-269, Berlin, Heidelberg, 2010. Springer-Verlag.

Questions and Answers

THANK YOU