SOLE-R, a Semantic and Linguistic Approach for Book Recommendations

Angel L. Garrido*, Maria Soledad Pera† and Sergio Ilarri*
*IIS Department, University of Zaragoza, Zaragoza, Spain.
Email: garrido@unizar.es, silarri@unizar.es
†Computer Science Department
Brigham Young Universty, Provo, Utah, USA
Email: mpera@cs.byu.edu

***Abstract*—Reading is a fundamental skill that each person needs to develop during early childhood and continue to enhance into adulthood. While children/teenagers depend on this skill to advance academically and become educated individuals, adults are expected to acquire a certain level of proficiency in reading so that they can engage in social/civic activities and successfully participate in the workforce. A step towards assisting individuals to become lifelong readers is to provide them adequate reading selections which can cultivate their intellectual and emotional growth. With that in mind, we have developed SOLE-R, a topic map-based tool that yields book recommendations. SOLE-R takes advantage of lexical and semantic resources to infer the likes/dislikes of a reader and thus is not restricted by the syntactic constraints imposed on existing recommenders. Furthermore, SOLE-R relies on publicly-accessible data on books to perform an in-depth analysis of the preferences of a reader that goes beyond book content or reading patterns explored by existing recommenders. We have verified the correctness of SOLE-R using a popular benchmark dataset. In addition, we have compared its performance with (state-of-the-art) recommendation strategies to further demonstrate the effectiveness of SOLE-R.**

***Keywords*-topic maps, recommendation systems, books**

I. Introduction

According to the National Institute of Child Health and Human Development, “reading is the single most important skill necessary for a happy, productive, and successful life”.[1] Unfortunately, a significant number of children/teenagers are below-average readers. The 2013 National Assessment of Educational Progress[2] reports that only 32% of American 4^{th} graders are proficient in reading. This finding is echoed by the 2013 report made by UNESCO Institute for Statistics which indicates that global illiterate population among youth ascends to 123.5 million[3]. It is imperative to encourage positive reading habits among children and teenagers due to the clear correlation between the academic performance of students and their reading ability [4]. In the aforementioned report, UNESCO also states that there are 773.5 million adults with rudimentary or no reading skills worldwide. Adults who cannot read to their children can negatively impact their children/teenagers’ early development. Given that reading affects both the intellectual and emotional growth of individuals, it is indispensable to provide adequate reading selections and encourage good reading habits among them. A step towards achieving this goal is to identify the right material for the right audience [20]. Finding *relevant items*, such as (non-)fiction books, however, can be challenging, given the huge amount of materials on diverse topics which are being published on a regular basis. A reader may turn to tools available at popular book-related websites, such as Amazon.com, to search for reading materials in various domains. These search tools, however, not only are inadequate for conducting personalized/contextual searches [5], but they also present users with an overwhelming number of items to choose from [5]. Recommendation systems, on the other hand, are the solution to the problem, since they can assist readers to cope with the *information overload* problem and minimize the *time* and *effort* imposed on discovering unknown items.

A number of recommenders in the reading domain have been developed over the past decades [16], which are employed by well-known commercial websites and social bookmarking sites. We have detected, however, many deficiencies in the design methodologies of these recommenders. Regarding the data required to make suggestions, current state-of-the-art methodologies (i) have not eradicated the *cold start* or *long-tail* problems [19] and (ii) are constrained by the requirement of large historical data on their users, which include, but are not limited to, personal tags and purchasing/accessing patterns, that might not be easily assembled. Moreover, these recommenders may not consider the users’ individual preferences, which is a major concern since suggesting materials that involve topics unappealing to the readers could diminish their interest in reading [2].

Given that both children and adults spend more time reading books[4], as opposed to other reading materials, we have focused our work on the recommendation of books and developed SOLE-R, a semantic, ontological, and linguistic enhanced recommender. SOLE-R takes advantage of natural language processing (NLP) tools and applies modern methodologies for knowledge management to provide

[1] http://www.ksl.com/?sid=15431484

[2] http://nationsreportcard.gov/reading_math_2013/#/student-groups

[3] http://goo.gl/08sbBT

[4] http://goo.gl/EmBx6h

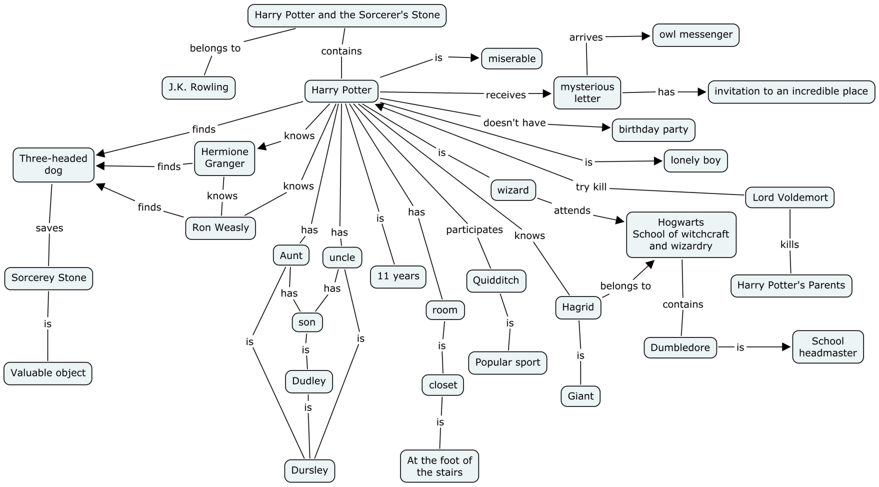

Figure 1. Example of a topic map for "Harry Potter and the Sorcerer's stone"

personalized book suggestions tailored to the preferences of individual users. Unlike its counterparts, SOLE-R examines the "meaning" of textual book metadata, such as content descriptions and reviews on books, considered during the recommendation process, as opposed to simply syntactically analysing the words in the texts, which translates into more suitable recommendations. There are already some semantic approaches as [18] but SOLE-R differs from them by the way the system generates abstractions of themes and subject areas from books and user profiles. As a representation of a conceptualization corresponds to the definition of an ontology [8], we can use techniques and methodologies for ontological engineering to model these representations and work with them. Such a representation makes it possible for SOLE-R to suggest new books of interest to a user in a more efficient manner, since it depends on concepts' associations, rather than simple words' associations. The major design goal of our recommendation strategy is to provide reading choices to users to motivate them to read, which could enrich their learning experience, and subsequently sharpen their critical thinking and analytical skills.

In the remainder of this paper, we discuss the design methodology of SOLE-R and present the results of the empirical study conducted to assess the correctness of our proposed recommender. Thereafter, we offer some concluding remarks and directions for future work.

II. Our proposed recommender

SOLE-R, our proposed recommender, adopts a topic map-based approach to make personalized suggestions suitable to each of its users. In the following sections, we first provide a brief overview of topic maps (see Section II-A). Thereafter, in Sections II-B to II-D, we detail the strategy adopted by SOLE-R to predict the degree to which a book unknown to a user matches his/her interests and preferences.

A. Topic Map Representations

A topic map is an specific type of concept map, i.e., a type of diagram, that shows relationships between concepts within a context [15]. An example of a topic map generated for the book "Harry Potter and the Sorcerer's Stone" by J.K. Rowling is shown in Figure 1. As previously stated, SOLE-R depends on topic maps to depict books and users, since topic maps, which are simple forms of ontologies, are very useful to represent knowledge.

B. Topic Maps of Books

The first step required by SOLE-R to make suggestions is to identify the content and other descriptive aspects of each book included in the set of books that have been bookmarked (i.e., read) by a user U, i.e., `list-Bt`. To build such a representation on books, SOLE-R adopts TM-Gen [7], which is a tool that extracts information from any number of texts and represents them in a topic map format [15].

To create `TM-B`, i.e., the topic map of each book B included in `list-Bt`, SOLE-R first examines a set of book descriptions on (the content of) B, i.e., `list-Sb`, in addition to a set of book reviews pertaining to B, i.e., `list-Rw`, using Freeling [3], an NLP tool, and WordNet [14], a well-known lexical database. It then proceeds to extract concepts and the corresponding relationships among them. Both content descriptions and reviews on books are obtained from reputable book-related websites, such as Amazon.com. The aforementioned textual representations on books differ on their contents, styles, and the type of language employed to write them. However, they capture book-related information from multiple perspectives and thus provide elements that are very useful in recommending a book to a user, which is the reason why SOLE-R examines both brief descriptions and reviews of books.

In building `TM-B`, SOLE-R needs to first generate a topic map, i.e., `TM-Sb`, for each description of B in `list-Sb`. Thereafter, these topic maps are merged into a single one, i.e., `TM-Sum`. To do this, we use SIM (Subject Identity Measure) [13], an existing approach that describes the relationships among two subjects or topics. We have chosen SIM because it yields good results for both recall and precision merging topic maps, and because it is language-independent.

As part of the topic map generation process, SOLE-R performs a semantic analysis of the topic map `TM-Sb`, and simplifies it if the system finds redundancies, incompatible associations, or ambiguities, using for this purpose a lexical database (i.e., WordNet) that contains semantic information of the words of a language. Similarly, SOLE-R identifies ambiguities in the topics. Thus, two or more topics can have the same meaning in a certain context but not in others. In this case, SOLE-R uses a disambiguation engine [11] taking into account associations and topics that are close to those concepts in the topic map hierarchy. If the process finds that they are synonyms, then it reduces them in the way described above.

SOLE-R proceeds to analyze each review in `list-Rw` to find relevant information on B, i.e., `Info-Rw`, which is used to enrich `TM-B`. As the language used in the reviews is usually much less formal than the one employed in book descriptions, it is more difficult to use parsers to extract information. For this reason, SOLE-R lemmatizes the texts in the reviews and extracts the most frequent keywords and Named Entities using the well-known TF-IDF algorithm [17]. These extracted keywords and Named Entities are incorporated into the topic map of B as new elements (either as topics or as relationships, using Freeling's morphological analyzer to classify them).

The set of topic maps `list-TM-B` includes all the enriched topic maps created for each of the books in `list-Bt`. The pseudo-code of the algorithm used by SOLE-R to generate `list-TM-B` is shown below.

```
/* Training books */
list-TM-B := empty-list;
FOR EACH b IN list-Bt DO
    list-Sb := Get_summaries(b);
    list-TM-Sb := empty-list;
    FOR EACH s IN list-Sb DO
        TM-Sb := Generate_topic_map(s);
        list-TM-Sb.add(TM-Sb);
    END FOR
    TM-Sum := Merge(list-TM-Sb);
    list-Rw := Get_reviews(b);
    Info-Rw := empty-list;
    FOR EACH rw IN list-Rw DO
        rw-inf := Analyze_review(rw);
        Info-Rw.add(rw-inf);
    END FOR
    TM-B := Enrich_TM(TM-Sum, Info-Rw);
    list-TM-B.add(TM-B);
END FOR
```

C. Topic Maps of Users

The next step in SOLE-R's recommendation process is to construct a profile of U which captures his/her interests/preferences, by examining the ratings that U has assigned to each book in `list-Bt`. In doing so, SOLE-R generates two different topic maps: `TM+`, which captures information about books favored by U, i.e., books that have been assigned a high rating by U, and `TM-`, which identifies information about books unappealing to U, i.e., low-rated books. SOLE-R generates `TM+` and `TM-` by merging (as discussed in Section II-B) the high-rated and low-rated books in `list-Bt`. The thresholds required to decide if a rating should be treated as "high" or "low" can be adjusted to suit the needs of each case. In the current implementation of SOLE-R, however, we set these thresholds to be 7 and 4, respectively. In other words, we treat books that have been given a rating of 7 or higher as "'high" and books with a rating of 4 or lower as "low".

Below is the pseudo-code of the algorithm applied by SOLE-R to yield the topic maps that capture the likes and dislikes of U.

```
/* Training user */
TM+ := NULL;
TM- := NULL;

list_+ := Get_best_rated(list-TM-B);
list_TM_- := empty-list;
FOR EACH b IN list_+ DO
    TM-B := Get_TM(b);
    list_TM_+.add(TM-B);
END FOR
TM+ := Merge(list_TM_+);

list_- := Get_worst_rated(list-TM-B);
list_TM_- := empty-list;
FOR EACH b IN list_- DO
    TM-B := Get_TM(b);
    list_TM_-.add(TM-B);
END FOR
TM- := Merge(list_TM_-);
```

D. Book Suggestions

The last step applied by SOLE-R in making suggestions involves predicting the degree to which U will like (or not) a new book `NB`. Using the approach described in Section II-B, SOLE-R generates `TM-NB`, the topic map of `NB`. Thereafter, SOLE-R evaluates the degree of similarity between `TM-B` and each of the topic maps that capture the likes/dislikes of U, i.e., `TM+` and `TM-`.

To calculate the similarity between topic maps, SOLE-R employs an algorithm we developed for SOLE-R that evaluates the resemblance between the topics of any two topic maps. This algorithm is based on two measures introduced in [12]: *lexical similarity* and *relational overlap*. While the first measure calculates the lexical overlap between strings, the second one quantifies the degree to which the relations of two concepts in an ontology match.

Using Equation 1, SOLE-R yields a single score for `NB` on a 1 to 10 range, where 1 (10, respectively) denotes that U finds `NB` unappealing (appealing, respectively).

$$Rate(NB) = Norm[(Sim(TM+) - Sim(TM-))] \quad (1)$$

where `Sim(TM+)` (`Sim(TM-)`, respectively) captures the degree of similarity between `TM-NB` and `TM+` (`TM-NB` and `TM-`, respectively), and $Norm$ is a function that maps the differences in similarity scores of `TM-NB` with respect to `TM+` and `TM-` from a [-1, 1] range to a [1, 10] range[5].

III. Experimental Results

In this section, we first introduce the dataset and framework we have considered to evaluate our proposed recommender. Thereafter, we discuss the results of the empirical studies conducted to assess the performance of SOLE-R, which we have compared with a number of baseline and state-of-the-art recommendation strategies.

A. Dataset

To evaluate the performance of SOLE-R, we turn to the BookCrossing dataset [22], which is a popular benchmark dataset commonly-used to assess the performance of book recommendation systems. BookCrossing, which we denote BX, contains information on 278,858 users and provides 1,149,780 ratings that have been assigned to 271,379 books.

B. Metric and Evaluation Strategy

To evaluate the performance of SOLE-R (and other recommendation strategies considered for comparison purpose) we apply the popular five-fold cross validation protocol. For each one of the five repetitions, 85% of the books rated by a user U in BX were used to model U's likes/dislikes (i.e., U_{train}) and the remaining 15% were used for the testing purpose (i.e., U_{test}).

In our empirical study, we quantify the performance of a recommender system R using the Root Mean Squared Error (RMSE), as shown in Equation 2, which is a de-facto metric for evaluating predictive recommendation systems, such as SOLE-R. Note that the $lower$ the RMSE computed for R is, the $better$ its performance, given its ability to correctly predict the degree to which a book will (or not) be of interest to a user. The RMSE scores reported in Section III-C are the $average$ of the RMSE scores computed for each of the five runs of the conducted experiments.

$$RMSE(R) = \frac{\sum_{U \in BX} \sqrt{\frac{\sum_{b \in U_{test}} |R_{U,b} - r_{U,b}|}{|U_{test}|}}}{|BX|} \quad (2)$$

where $R_{U,b}$ denotes the rating $predicted$ by R for a book b ($\in U_{test}$) given the corresponding user U and $r_{U,b}$ is the $actual$ rating given to b by U.

C. Empirical Study

As shown in Figure 2, the RMSE score generated using SOLE-R is 1.25. To further assess the effectiveness of the recommendation strategy of SOLE-R, we compare its performance, in terms of RMSE, with a number of baseline recommenders: ALSWR (alternating least squares with regularization) [21], SGD [6], SVD++ [10], and Bias-SVD [10]. We also compare SOLE-R with a number of state-of-the-art recommenders: BMF-B [9], BMF-R [9], fLDA [1], RLMF [1], and uLDA [1]. Note that, due to space constraints, we are unable to provide details on these recommenders, so please refer to the references for an in-depth discussion on each of their methodologies.

Based on the conducted empirical study, we have verified that SOLE-R outperforms the aforementioned recommendation strategies, given the statistically significant ($p < 0.001$) difference in RMSE yielded by SOLE-R with respect to each of its counterparts.

IV. Conclusion and Future Work

In this paper, we have presented a new method to face the problem of providing adequate reading selections to individuals. The algorithm combines semantic and ontological techniques with NLP tools and lexical resources to made book recommendations suitable to the preferences/interests of each individual user.

Our main contributions are:

- Testing semantic and NLP techniques in the field of recommendation systems.
- Treating each element to be recommended as a knowledge network, and thus being able to take advantage of ontological methods to compare it with user preferences, mapped in the same way.
- Improving the results achieved by well-known techniques using a new algorithm supported by NLP tools, ontologies, and semantic methods.

We have conducted an empirical study using the BookCrossing dataset. The computed results verified not only the correctness and effectiveness of SOLE-R, our proposed recommender, but also its superiority over a number of baseline and state-of-the-art recommendation strategies, which are based on methodologies commonly-used for performing the recommendation task.

As previously stated, the main goal of our recommender is to locate the right material for the right audience. In doing so, we aim not only to promote good reading habits among readers of all ages, but also at facilitating and encouraging their learning. While we have empirically verified the correctness of SOLE-R, we would like to further extend

[5] Note that even though the current implementation of SOLE-R predicts the degree of interest of a user in a given book on a [1, 10] range, this can easily be adapted to comply with other common rating scales.

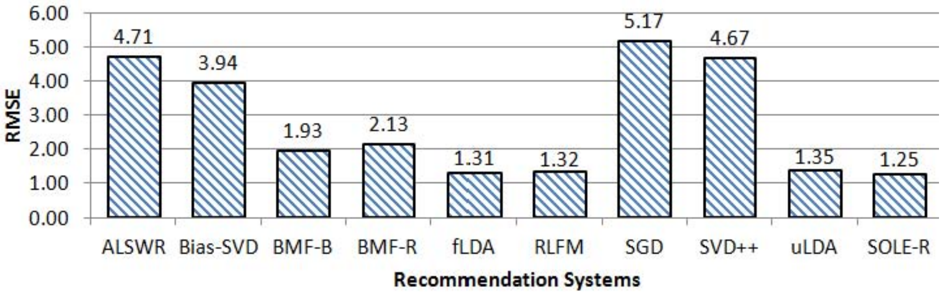

Figure 2. Performance evaluation of SOLE-R and a number of recommendation strategies considered for comparison purposes

its evaluations by conducting in-depth user-studies on K-12 and advanced readers, which would allow us to quantify the degree of influence of SOLE-R towards encouraging reading/learning.

Acknowledgment

The authors would like to thank Maria G. Buey, Sandra Escudero and Alvaro Peiro their collaboration. This research work has been supported by the CICYT project TIN2010-21387-C02-02 and DGA-FSE.

References

[1] D. Agarwal and B. Chen. *fLDA: Matrix Factorization through Latent Dirichlet Allocation*. ACM WSDM, pp. 91-100, 2010.

[2] R. Allington and E. Gabriel. *Every Child, Every Day*. Reading: The Core Skill, 69(6):10-15, 2012.

[3] X. Carreras, I. Chao, L. Padró, and M. Padró. *FreeLing: An Open-Source Suite of Language Analyzers*. LREC pp.239-242, 2004.

[4] S. Chu and S. Tse and E. Loh and K. Chow. *Collaborative Inquiry Project-based Learning: Effects on Reading Ability and Interests*. Library and Information Science Research, 33:236-243, 2011.

[5] G. De Francisci Morales and A. Gionis and C. Lucchese. *From Chatter to Headlines: Harnessing the Real-Time Web for Personalized News Recommendation*. ACM WSDM, pp. 153-162, 2012.

[6] S. Funk. *Stochastic Gradient Descend*. Available at: http://sifter.org/ simon/journal/20061211.html, 2006.

[7] A. Garrido, M. Buey, S. Escudero, S. Ilarri, E. Mena, and S. Silveira. *TM-Gen: A Topic Map Generator from Text Documents*. IEEE ICTAI, pp.735-740, 2013.

[8] T. Gruber. *A Translation Approach to Portable Ontology Specifications*. Knowledge Acquisition, 5(2):199-220. Academic Press Ltd. 1993.

[9] R. Kannan, M. Ishteva, and H. Park. *Bounded Matrix Factorization for Recommender Systems*. Knowledge and Information Systems, p. 1-21, 2013.

[10] Y. Koren. *Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model*. ACM SIGKDD, pp. 426-434, 2008.

[11] R. Navigli. *Word Sense Disambiguation: A Survey*, ACM Computing Surveys, 41(2), 2009.

[12] A. Maedche and S. Staab. *Measuring Similarity Between Ontologies*. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Springer, pp. 251-263, 2002.

[13] L. Maicher and H. Witschel. *Merging of Distributed Topic Maps based on the Subject Identity Measure (SIM) approach*. Berliner XML-Tage, 4:301-307, Tolksdorf, Eckstein, 2004.

[14] G. Miller. *WordNet: A Lexical Database for English*. Communications of ACM, 38(11):39-41, 1995.

[15] S. Pepper and G. Moore. *XML T Maps (XTM) 1.0 - TopicMaps. Org Specification*. TopicMaps.Org Authoring Group, http://www.topicmaps.org/xtm, 2001.

[16] F. Ricci, L. Rokach, B. Shapira, and P. Kantor. *Recommender Systems Handbook*, Springer, 2011.

[17] G. Salton and C. Buckley. *Term-weighting Approaches in Automatic Text Retrieval*. Information Processing and Management, 24(5):513-523, 1988.

[18] G. Semeraro and P. Basile and M. de Gemmis and P. Lops. *Content-Based Recommendation Services for Personalized Digital Libraries*. Digital Libraries: Research and Development, pp. 77-86. Springer, 2007.

[19] M. Sun and F. Li and J. Lee and K. Zhou and G. Lebanon and H. Zha. *Learning Multiple-question Decision Trees for Cold-start Recommendation*. ACM WSDM, pp. 45-454, 2013.

[20] S. Vanneman. *Keep Them Reading*. School Library Monthly, 27(3):21-22,2010.

[21] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. *Large-scale Parallel Collaborative Filtering for the Netflix Prize*. Algorithmic Aspects in Information and Management. LNCS 5034:337-348, 2008.

[22] C. Ziegler, S. McNee, J. Konstan, and G. Lausen. *Improving Recommendation Lists Through Topic Diversification*. ACM WWW, pp. 22-32, 2005.