IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2017, accepted April 14, 2017, date of publication April 26, 2017, date of current version June 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2698142

Semantically Enhanced Medical Information Retrieval System: A Tensor Factorization Based Approach

HAOLIN WANG[1,2,3], QINGPENG ZHANG[3,4], (Member, IEEE), AND JIAHU YUAN[1,2]

[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong
[4]Shenzhen Research Institute of City University of Hong Kong, Shenzhen 518057, China

Corresponding author: Qingpeng Zhang (qingpeng.zhang@cityu.edu.hk)

This work was supported in part by the National Natural Science Foundation of China under Grant 71402157 and Grant 71672163, in part by the Guangdong Provincial Natural Science Foundation under Grant 2014A030313753, and in part by the Theme-Based Research Scheme of the Research Grants Council of Hong Kong under Grant T32-102/14N.

ABSTRACT Medical information retrieval plays an increasingly important role to help physicians and domain experts to better access medical-related knowledge and information, and support decision making. Integrating the medical knowledge bases has the potential to improve the information retrieval performance through incorporating medical domain knowledge for relevance assessment. However, this is not a trivial task due to the challenges to effectively utilize the domain knowledge in the medical knowledge bases. In this paper, we proposed a novel medical information retrieval system with a two-stage query expansion strategy, which is able to effectively model and incorporate the latent semantic associations to improve the performance. This system consists of two parts. First, we applied a heuristic approach to enhance the widely used pseudo relevance feedback method for more effective query expansion, through iteratively expanding the queries to boost the similarity score between queries and documents. Second, to improve the retrieval performance with structured knowledge bases, we presented a latent semantic relevance model based on tensor factorization to identify semantic association patterns under sparse settings. These identified patterns are then used as inference paths to trigger knowledge-based query expansion in medical information retrieval. Experiments with the TREC CDS 2014 data set: 1) showed that the performance of the proposed system is significantly better than the baseline system and the systems reported in TREC CDS 2014 conference, and is comparable with the state-of-the-art systems and 2) demonstrated the capability of tensor-based semantic enrichment methods for medical information retrieval tasks.

INDEX TERMS Information retrieval, tensor factorization, knowledge based systems.

I. INTRODUCTION

With the exponential growth of medical related information, the retrieval of high quality results is becoming more critical. For medical applications like the Clinical Decision Support System (CDSS), an effective and reliable information retrieval (IR) system is the basis to provide scientific evidences to support the clinical decision making, facilitate the translation of latest research outcomes into practice and improve the quality of health care [1]. The challenges are from (a) the inherent complexity of medical languages such as obscure medical terminologies and ambiguous abbreviations, and (b) the associated variety of information needs from different types of users, such as patients and physicians [2]. The complexity and ambiguity of medical languages result in vocabulary mismatch between queries and documents, which makes the conventional keyword-based IR methods often ineffective in medical IR tasks. Semantic knowledge bases and concept mapping techniques are with the potential to solve this problem through annotating and analyzing the information with a common medical terminology and semantic relations including the classifications, term dependencies, hierarchies, etc.

Integrating the semantic relations is able to access and use the rich information of domain knowledge to enable accurate inferences for varying information needs. Ontologies and associated semantic information are useful resources to

extract structured knowledge for IR. For instance, a user submits a query to search for the information of fever treatment. There is a document introducing ''aspirin'', a medication used to treat fever. Human experts may judge that this document is relevant to the user's query. However, this relevance could not be automatically identified without the knowledge bases which contain the association in the form of a triple (''aspirin'', ''may_treat'', ''fever'').

Studies also showed that the proper incorporation of semantic information in knowledge bases is critical to the performance of medical IR [3]. One of the most effective approaches is knowledge-based query expansion, which is a well-known method to bridge the gap between query terms and actual user information needs [4]. With the expanded set of terms, there is a higher probability to identify and retrieve relevant documents that do not contain the terms in the original query. In addition, query expansion is flexible to be integrated with existing IR systems.

However, there are a number of problems that need to be addressed when we incorporate the semantic information in knowledge bases: (a) Most medical knowledge databases suffer from the incompleteness and lack of reasoning capability over relations [5]; (b) The noisy samples and the inaccuracy of concept mapping set higher requirement for the model to achieve good performance. For instance, a study showed that the precision of concept mapping through MetaMap is only 71.8% [6]; (c) Similar to many real-world datasets, the observed data is generally sparse because of the incompleteness of knowledge bases and the limited annotated samples. Therefore, there is a crucial need to develop a systematic methodology to address the aforementioned problems, and effectively incorporate the semantic knowledge bases for medical IR.

To address these challenges, we developed a semantically enhanced medical IR system based on a two-stage query expansion strategy. The system first expanded the original query through incorporating new terms from initially retrieved documents. Then, we proposed a latent semantic relevance model for the system to capture the relevant concepts and semantic relations between concepts through extracting semantic triples from knowledge bases. The system built the third-order tensor to represent the ternary relations, and adapted tensor factorization methods to estimate the latent features of the semantic association triples. Advantages of the proposed tensor factorization based latent feature model are summarized as follows: (a) the incompleteness problem can be formed as filling in missing entries in relational datasets. Tensor factorization is widely used for this kind of problems such as link prediction [7]. (b) Multi-way analysis is more robust to noise over two-way matrix analysis benefiting from the power of multi-liner algebra [8]. (c) Tensor factorization provides the unique capability to work well under sparse settings such as recommender systems [9]–[11]. Tensor factorizations have been demonstrated to be an effective way to resolve the sparsity problem by breaking the independence of multiple interaction parameters [12]. Based on the proposed latent semantic relevance model, we identified effective semantic patterns to trigger knowledge-based query expansion. Our system is able to incorporate medical knowledge bases to infer the actual user needs and improve the performance of medical IR tasks. Experiments with the TREC CDS 2014 dataset demonstrate that the performance of the proposed system is significantly better than the baseline and the best results reported in TREC CDS 2014 conference, and is comparable with state-of-the-art approaches on this retrieval task.

The rest of this paper is organized as follows. Section II summarizes related work on semantic IR and tensor factorization based applications. Section III introduces notations and preliminaries. The proposed medical IR system is presented in Section IV. Experiment results are evaluated in Section V. We conclude the paper in Section VI with discussions of future work.

II. RELATED WORK

In this section, we review the literatures related to this work from two perspectives: (a) knowledge-based medical information retrieval, and (b) the applications of tensor factorization in data mining.

A. KNOWLEDGE-BASED MEDICAL INFORMATION RETRIEVAL

Medical information retrieval aims to discover the scientific evidences to support decision making with medical domain knowledge. Knowledge bases developed by domain experts are able to enhance the understanding of free text with domain knowledge. Structured knowledge bases represent the information as a set of knowledge graphs, which consist of the entities (nodes) and the relations (edges) between them. This kind of knowledge representation has a long history in logic and artificial intelligence. More recently, it is used in the Semantic Web Community to create a ''web of data'' which is processable and comprehensible by computer programs [13]. In particular, the Unified Medical Language System (UMLS) is a widely used knowledge base in medical domain. UMLS is a repository of biomedical terms, concepts, and the relations between them from multiple resources [14].

Several studies explored the integration of knowledge bases to improve medical IR performance. Liu and Chu proposed a knowledge-based query expansion method which appends the original query with additional terms that are specifically relevant to the query's scenarios such as diagnosis and treatment of diseases [15]. Martinez *et al.* presented an automatic query expansion method based on random walks over the UMLS semantic knowledge base, and showed that query expansion with terms beyond synonymy is an effective approach to identify the similarity between the query and documents [16]. Otegi *et al.* performed both query expansion and document expansion using a lexical database on IR tasks for question answering, and showed that their methods are complementary with pseudo-relevance feedback [17]. Sfakianaki *et al.* proposed a natural language

processing framework to automatically transform a clinical research question to a query that contains only terms of biomedical ontologies. Their research demonstrated the capability of biomedical ontologies and entity annotation algorithms to bridge the gap between clinical questions in natural language and biomedical literature [18]. Mao *et al.* proposed a new medical IR system enhanced by manually assigned subject terms (Medical Subject Headings, MeSH). The proposed system constructs generative concept models to capture the associations between queries and documents [19]. Koopman *et al.* proposed a medical IR system that integrates structured knowledge resources, statistical information retrieval methods, and the semantic inference in a unified framework for large-scale IR applications [20]. In addition, indexing and retrieval software packages (i.e. Lucene and Indri) are widely adopted to build large-scale IR applications [21], [22]. In this paper, we integrated Solr, a search platform built on Lucence, as the base component to build the medical IR system [23].

In particular, providing access to relevant biomedical literature in a clinical setting has the potential to enable and support the evidence-based medicine. In order to encourage the research in this field, the Text REtrieval Conference (TREC) has been hosting the Clinical Decision Support (CDS) track since 2014 [24]. TREC CDS track released the dataset and requested participants to develop effective medical IR systems to provide relevant medical documents to clinicians to improve their decision-making in diagnosing, treating, and testing patients. The integration of medical knowledge bases and query expansion methods are widely used by the participants of this track. However, the performance of knowledge-based approaches was not satisfactory. The organizers pointed out that the poor performance of existing approaches could be resolved with available training data to tune the parameters [25]. In this paper, the available TREC CDS data is leveraged to explore a new supervised learning approach to enhance the performance of knowledge-based medical IR.

B. TENSOR FACTORIZATION FOR DATA MINING

Tensors are multidimensional arrays to describe the linear relations between objects. Tensors provide a natural framework for representing and solving problems in a wide range of areas. Tensor factorization is the higher-order extensions of matrix factorization, which is able to capture the latent patterns in multi-way datasets [26]. Tensor factorization (or tensor decomposition) has been successfully applied to a rich set of data mining and machine learning problems including computer vision, link prediction, recommender systems etc. Compared with matrix decomposition approaches, tensor factorization explicitly exploits the multi-way structure which will be lost when collapsing the tensor to matrices. The most well-known decomposition methods are CANDECOMP/PARAFAC (CP) decomposition and Tucker decomposition, which can be treated as the high order extension of matrix singular value decomposition (SVD) and principal component analysis (PCA) [26]. Shashua and Hazan introduced the non-negative tensor factorization algorithm for the applications in computer vision such as sparse image encoding [27]. Dunlavy *et al.* illustrated the three-dimensional tensor model are effective for temporal link prediction using CP decomposition [28]. Tensor based methods have been widely used in recommender systems through identifying the underlying similarity and association between objects with factorization techniques [29]. In Semantic Web, Franz *et al.* proposed a three-dimensional tensor based approach for faceted authority ranking in the context of knowledge bases, and obtained better performance as compared with conventional graph-based ranking methods [30]. Chang *et al.* proposed a tensor factorization approach for knowledge based embedding to discover new relations missing in knowledge bases [31]. Nakatsuji *et al.* developed a new semantic sensitive tensor factorization method to incorporate semantics for recommender systems, and achieved a higher accuracy compared with other tensor-based methods without semantic information in knowledge bases [32]. Different from the aforementioned studies, we propose a novel method to capture the semantic relevance patterns based on tensor factorization to improve the performance of medical IR.

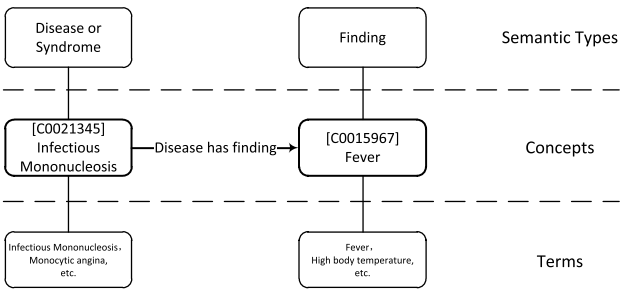


FIGURE 1. An example of the semantic information and relations extracted from UMLS.

III. NOTATIONS AND PRELIMINARIES

To develop the knowledge-based medical IR system, we incorporate the domain-specific information extracted from UMLS, a widely used knowledge base in medical domain. In the UMLS, synonymous terms are clustered into concept, and concepts are linked to other concepts in the semantic network. In addition, concepts are broadly categorized by means of semantic types [14]. For instance, the semantic information extracted for two concepts ''fever'' and ''infectious mononucleosis'' is shown in Figure 1. They are assigned with semantic types ''finding'' and ''disease or syndrome'' respectively, and connected with the semantic relation ''disease has finding'' in the UMLS semantic network. The terms such as ''high body temperature'' and ''fever'', which have the same meaning, are grouped into a unified concept with concept identifier ''C0015967''.

Let χ denote a tensor, which is the higher-order generalization of vector and matrix. The order (also known as way or mode) of a tensor is the number of dimensions. Vectors are first-order tensors, which are denoted by boldface

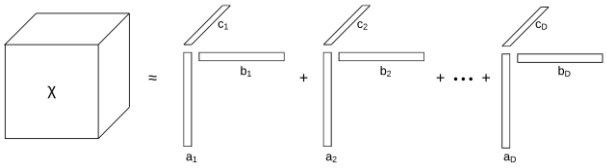


FIGURE 2. Illustration of CP Decomposition.

lowercase letters, e.g., $\boldsymbol{a}$. Matrices are second-order tensors, which are denoted by boldface capital letters, e.g., $\mathbf{A}$. An entry of a vector is a_i and an entry of a third-order tensor is x_{ijk}. The terms "factorization" and "decomposition" refer to the same process to factorize the tensor to generate decomposed components. There are multiple decomposition methods such as Tucker decomposition and CP decomposition. The Tucker decomposition decomposes a tensor into a core tensor that multiplied by a matrix along each mode. The CP decomposition decomposes a tensor into the sum of component rank-one tensors as shown as follows [26].

$$\boldsymbol{\chi} \approx \sum_{r=1}^{D} \boldsymbol{a_r} \circ \boldsymbol{b_r} \circ \boldsymbol{c_r} \tag{3.1}$$

The symbol "$\circ$" represents the outer product of the vectors. D represents the number of components in this CP model.

The smallest number of components in an exact CP decomposition, which means equality holds in (3.1), is defined as the tensor rank. The definition of tensor rank is similar to matrix rank but they have quite different properties. Particularly, the tensor rank cannot be calculated by any known algorithm. For a general third-order tensor $\boldsymbol{\chi} \in R^{I \times J \times K}$, a weak upper bound on its maximum rank is known as [33].

$$rank\,(\boldsymbol{\chi}) \leq min(IJ, JK, IK) \tag{3.2}$$

In practice, the number of components in the CP decomposition, as D in (3.1), is usually determined by fitting various rank-$\boldsymbol{D}$ CP models, which will be further discussed in Section IV.B.2.

The CP decomposition can be considered as a special case of the Tucker decomposition, with the advantage of uniqueness [34]. For the rest of this paper, CP decomposition is adopted as the method for tensor factorization, and we use a simple third-order tensor as an example to explain how to use CP decomposition to extract the latent feature representations for objects encoded in the tensor. A typical use of the third-order ($R^{m \times n \times l}$) tensor is to model the relational data such as the interactions of objects A, B and C as shown in Figure 3. An entry x_{ijk} in the tensor denotes the interaction of the triple (A_i, B_j, C_k).

In this example, each entry in the tensor model is the inner product of three decomposed latent feature vectors $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{iD})$, $\boldsymbol{b}_j = (b_{j1}, b_{j2}, \ldots, b_{jD})$ and $\boldsymbol{c}_k = (c_{k1}, c_{k2}, \ldots, c_{kD})$, in the form of

$$x_{ijk} \approx \sum_{r=1}^{D} a_{ir} b_{jr} c_{kr} \tag{3.3}$$

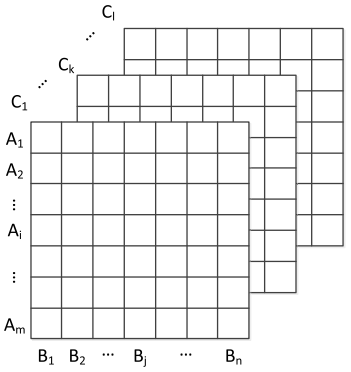


FIGURE 3. The 3-order tensor model for relational data.

Thus, the latent feature representation is created for objects encoded in the tensor. These vectors are called the latent feature representation because they could not be observed explicitly, and they define both the representation and relations between these objects in the latent feature space. Once the decomposed representation is found, the interaction of any triple $(\mathrm{A_i}, \mathrm{B_j}, \mathrm{C_k})$ can be recovered with the equation (3.3).

The rest part of this section is the preliminaries for the alternating least squares method for the CP decomposition. The factor matrices are the combinations of the decomposed latent feature vectors, e.g. $\mathbf{A} = [\boldsymbol{a_1}\ \boldsymbol{a_2} \ldots \boldsymbol{a_D}]$ and likewise for $\mathbf{B}$ and $\mathbf{C}$. Then the CP decomposition can be expressed as $\boldsymbol{\chi}'$, which is an approximation of the original tensor $\boldsymbol{\chi}$.

$$\boldsymbol{\chi} \approx \boldsymbol{\chi}' = [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!] \tag{3.4}$$

To reduce the complexity of tensor decomposition problem, the third-order tensor can be unfolded into matrices through one of the three modes represented by $\boldsymbol{\chi}_{(1)}$, $\boldsymbol{\chi}_{(2)}$ and $\boldsymbol{\chi}_{(3)}$. For example, one of the matricized forms of the tensor is as follows.

$$\boldsymbol{\chi}_{(1)} \approx \boldsymbol{\chi}'_{(1)} = \boldsymbol{A}\,(\boldsymbol{C} \odot \boldsymbol{B})^{\boldsymbol{T}} \tag{3.5}$$

And it has the property that

$$\boldsymbol{A} = \boldsymbol{\chi}'_{(1)}\,(\boldsymbol{C} \odot \boldsymbol{B}) \left(\left(\boldsymbol{C^T C}\right) \times \left(\boldsymbol{B^T B}\right)\right)^{\dagger} \tag{3.6}$$

The symbol "$\odot$" denotes the *Khatri-Rao product* and "$\dagger$" denotes the *Moore-Penrose pseudoinverse*. See [26] for details.

IV. THE PROPOSED MEDICAL IR SYSTEM

We developed a semantically enhanced medical IR system, which has a two-stage query expansion strategy (as shown in Figure 4) to integrate the pseudo relevance feedback and the knowledge-based query expansion to improve the performance of retrieving relevant documents for queries. In general, relevance feedback is to statistically discover the implicit relevant information to improve the recall of retrieved documents, while knowledge-based query expansion is to infer a user's actual needs with domain knowledge

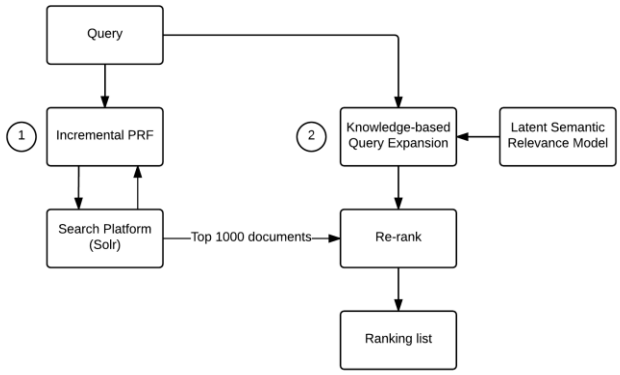


FIGURE 4. The framework of the proposed medical IR system.

to improve the precision. First, we proposed the incremental pseudo relevance feedback (incremental PRF) approach for query expansion to obtain the initial ranking list of retrieved documents. Second, we developed an enhanced knowledge-based query expansion method with a novel latent semantic relevance model. The proposed method will re-rank the documents retrieved by the incremental PRF in the first stage. For instance, when a user input a query, the system will first (a) use incremental PRF to expand the query and obtain the top 1000 relevant documents based on the search platform (Solr). Then, the system will (b) use the proposed latent semantic relevance model to expand the query based on medical knowledge bases. In the end, the system will (c) use the expanded query based on knowledge to re-rank the 1000 relevant documents as the final retrieval result.

A. INCREMENTAL PSEUDO RELEVANCE FEEDBACK

Relevance feedback methods are widely used to reformulate the original query using expansion features from the retrieved relevant documents. The popular pseudo-relevance feedback method assumes that top retrieved documents are relevant to the query so that the terms from these documents can be used for expansion [4]. Usually a fixed setting is used across different queries, but it's not optimal because of the variety of queries and feedback documents. Improper expansion may cause query drift, and proper expansion could identify more relevant documents. The effectiveness of expansion methods can be evaluated directly from their impact on the IR result [35]. Therefore, an optimal approach is to extract proper and effective expansion terms from top-ranked retrieved documents for each query. In order to select proper expansion terms, we proposed a heuristic approach, which iteratively uses the scores returned by Solr retrieval system as an indicator to boost the similarity scores between queries and documents.

The Rapid Automatic Keyword Extraction (RAKE) is applied to extract keywords from top retrieved documents [36]. RAKE is an extremely efficient method, which can operate on individual documents and able to extract phrases instead of a single keyword. Using this algorithm, related keywords from potential related documents are exploited to improve the recall of the medical IR system. The incremental strategy is applied to evaluate the proper set of expansion terms iteratively and control the expansion process with a threshold to reduce the risk of query drift.

Algorithm 1 Incremental PRF

Input: *query*
incremental ratio α
Output: *a list of ranked documents for the input query*

1: Input *query* to the retrieval system to get the initial list of ranked documents
2: Calculate the average ranking *score* of the top m documents in the ranking list and set
$threshold = score \times \alpha, k = 0$
3: Repeat
4: Increase k by 1
5: Extract n keywords from each of the top k documents in the current ranking list
6: Update the *query* by adding the extracted n keywords into the original *query*
7: Input the updated *query* to the retrieval system
8: Until the average ranking *score* of the top m documents $\geq$ *threshold*,
9: Return the current list of ranked documents as the output.

B. LATENT SEMANTIC RELEVANCE MODEL

After we performed the incremental PRF, the system will obtain a list of top 1000 documents related to the query. In this section, we introduce the knowledge-based query expansion with a latent semantic relevance model, which is able to provide the optimal expansion paths under sparse settings, and we use the expanded query to re-rank the 1000 documents to obtain the final results.

1) TENSOR BASED SEMANTIC ASSOCIATION REPRESENTATION

To incorporate the UMLS semantic network for knowledge-based query expansion, the query terms are mapped to UMLS concepts first, and then the related concepts in the semantic network could be selected as expansion concepts. However, the semantic network covers a wide range of related concepts, some of which may be useless or even harmful for the retrieval of relevant documents. Intuitively, domain-specific semantic types (i.e. "disease or syndrome" and "sign or symptom") and semantic relations (i.e. "may treat" and "disease has finding") would play more important roles in medical IR through facilitating the relevance assessment between queries and documents. To effectively identify query concepts to be expanded, and the expansion paths through different semantic relations, we deploy a data-driven approach by extracting the semantic associations between queries and relevant/irrelevant documents. For instance, there is a query containing the concept "fever." Among all labeled

documents, there are 57 documents containing concepts of one or more drugs (i.e. ''aspirin'') for fever treatment. 50 out of these 57 documents are labeled as relevant to the query, while the other 7 are labeled as irrelevant. In this example, the semantic concept triples such as (''aspirin'', ''may_treat'', ''fever'') could be utilized to improve the performance by adding the expansion concepts ''aspirin'' for query concept ''fever''.

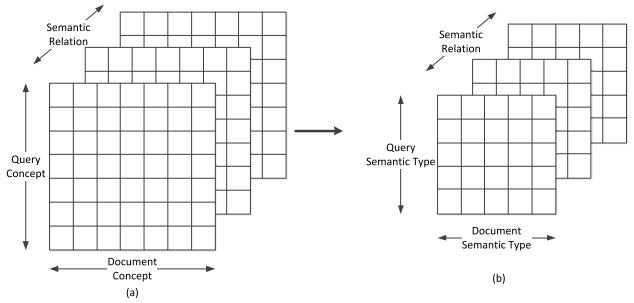


FIGURE 5. Example of tensor encoding of semantic association triples.

For medical IR tasks with large-scale dataset and a small number of labeled documents, it is unrealistic to capture every possible semantic association to support query expansion. In addition, the data could be very sparse given the large number of possible semantic associations. In UMLS, the semantic types (i.e. ''pharmacologic substance''), which represent a high level generalizations of semantic concepts (i.e. ''aspirin''), could be utilized to reduce the complexity without losing substantial details for specific tasks [37], [38]. For example, the semantic association triple for the aforementioned triple (''aspirin'', ''may_treat'', ''fever'') is (''pharmacologic substance'', ''may treat'', ''finding''). Therefore, in this paper, we extract the semantic association triples based on the semantic types (as shown in Figure 5). Further, as incorporating knowledge bases has the problems of incompleteness, noise and sparsity, we presented the latent semantic relevance model based on tensor factorization to (a) filter out the noise of observed semantic type triples, and (b) estimate the existence of unobserved triples for relevance assessment. Then a ranked list of all possible semantic type triples (including observed and unobserved triples) will be generated to serve as the candidate expansion paths for the corresponding query. Section IV.B.3 will explain the expansion process in details.

Benefiting from its high dimension, tensor provides the natural representation for heterogeneous features such as the dependencies between concept pairs with different semantic relations (Figure 5) in the form of (semantic type in the query, semantic relation, semantic type in the document). Different semantic relations have different impact on the relevance assessment, and multiple semantic relations make the model a natural third-order tensor.

We define the notations as query semantic types $\{q_i\}$, semantic relation $\{s_k\}$, and document semantic types $\{d_k\}$. We performed CP decomposition with D components so that each element in the tensor is calculated as the inner product of the three latent feature vectors $\boldsymbol{q}_i, \boldsymbol{s}_i$ and $\boldsymbol{d}_k$. For each triple (q_i, s_j, d_k), the value $f(q_i, s_j, d_k) = x_{ijk}$ is assigned to reflect the association of the three elements in the triple.

$$f(q_i, s_j, d_k) = \sum_{r=1}^{D} q_{ir} s_{jr} d_{kr} \tag{4.1}$$

x_{ijk} is then used to assess the semantic relevance of the corresponding queries and documents. Initially, we assign the value for observed triples based on the association rules to infer relevance based on the occurrence of each triple independently. Thus, we calculate the proportion of extracted triples in relevant documents to infer the conditional probability of relevance given the occurrence of this triple as below.

$$f_0(q_i, s_j, d_k) = \frac{C_r(q_i, s_j, d_k)}{C_f(q_i, s_j, d_k)} \tag{4.2}$$

$C_f(q_i, s_j, d_k)$ is the count of triple (q_i, s_j, d_k) of the full training set, and $C_r(q_i, s_j, d_k)$ is the count of observed triple $\left(q_i, s_j, d_k\right)$ of relevant training samples. The value for each triple is in the range of 0 to 1. The proper rank-D decomposition and then restoring the original tensor is an approach to eliminating part of the estimation variance and noise [39].

2) TENSOR FACTORIZATION

Estimation of the existence of unobserved triples for relevance assessment is based on the aggregation of possible semantic relations between concepts, which is inherently similar to the link prediction of semantic networks. Particularly, Alternating Least Squares method for CP decomposition (CP-ALS) yields a good performance for predicting semantic relations in a subset of UMLS [40]. Therefore, we adopted CP-ALS method for tensor factorization in our system.

For CP-ALS method, the rank of tensor is an input parameter to be determined. The choice of the rank for CP decomposition is still an open problem that cannot be solved by any known methods. In practice, the solution is to generate CP decompositions with different numbers of components until the fitting rate reaches a certain threshold. However, simply fitting the original tensor could not determine the optimal rank in this problem due to the existence of missing entries. A standard practice is to impute the missing value, and then use the factorization to re-impute the missing entries [41]. The optimal estimation of the rank will then be determined based on the fitting of the known entries in the original tensor. In this paper, we adopted a similar approach based on CP-ALS method and a heuristic optimization algorithm. First, the mean value of known entries is used as the initial value for estimation of missing entries. Second, we incorporated a weighted error function [41] for the rank estimation.

$$f_w(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left\{ w_{x_{ijk}} (x_{ijk} - x'_{ijk}) \right\}^2,$$

where

$$w_{x_{ijk}} = \begin{cases} 1, & \text{if } x_{ijk} \text{ is a known entry} \\ 0, & \text{if } x_{ijk} \text{ is a missing entry} \end{cases} \tag{4.3}$$

x'_{ijk} is the entry value of χ', which is recovered with the equation (3.3) after decomposition. The modified CP-ALS algorithm with a weighted error function has been adopted to provide the high quality estimation of latent features as follows.

Algorithm 2 Weighted Fitting for CP-ALS Decomposition for Third-Order Tensor

Input: tensor χ
minimal rank r_{min}
maximal rank r_{max}
error ratio ε
Output: *decomposed components*

1: $r = r_{min}$
2: Initialize A, B randomly
3: Repeat
4: Repeat
5: $\boldsymbol{C} = \boldsymbol{\chi}_{(3)}(\boldsymbol{B} \odot \boldsymbol{A})((\boldsymbol{B}^T\boldsymbol{B}) \times (\boldsymbol{A}^T\boldsymbol{A}))^{\dagger}$
6: $\boldsymbol{B} = \boldsymbol{\chi}_{(2)}(\boldsymbol{C} \odot \boldsymbol{A})((\boldsymbol{C}^T\boldsymbol{C}) \times (\boldsymbol{A}^T\boldsymbol{A}))^{\dagger}$
7: $\boldsymbol{A} = \boldsymbol{\chi}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B})((\boldsymbol{C}^T\boldsymbol{C}) \times (\boldsymbol{B}^T\boldsymbol{B}))^{\dagger}$
8: Recover the tensor $\chi' = [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]$
9: Calculate $f_w(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \sum_i^I \sum_j^J \sum_k^K \{w_{x_{ijk}}(x_{ijk} - x'_{ijk})\}^2$
10: Until $f_w(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ ceases to improve or maximum iterations exhausted
11: Increase r
12: Until $\frac{f_w(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})}{\|\chi\|} \leq \varepsilon$, OR $r = r_{max}$

3) KNOWLEDGE-BASED QUERY EXPANSION

The aforementioned tensor factorization method creates a latent feature vector for each semantic type and each semantic relation. The factorization of high order tensor usually has the problem of high computation cost. In our system, the factorization only need to be calculated once to estimate the score of the semantic association triples. Therefore, it will not affect the efficiency for the information retrieval task for any specific query. For both proposed query expansion methods, the complexity increases linearly corresponding to the iteration thresholds. The score for the semantic association triples including unobserved ones can then be calculated by taking the inner product of the corresponding latent feature vectors. These semantic association triples provide the possible expansion paths with semantic relation and semantic type constraints for the concepts in the semantic network. For each query, top-ranked triples which are applicable for the concepts of this query will be used as the expansion paths to identify expansion concepts. The associated terms of these expansion concepts will be added to the original queries to form the expanded queries. For example, a ranked list of triples is shown in Table 1. For the query containing the concept "fever", first we retrieve all associated concepts and their relations from the knowledge base. Then the triple ("Finding", "may_treat", "Organic Chemical")

TABLE 1. The example of ranked semantic association triples obtained from CP decomposition.

Semantic Association Triple	Score
…	…
("Finding", "may_treat", "Organic Chemical")	0.760
("Finding", "disease_has_finding", "Disease or Syndrome")	0.486
…	…

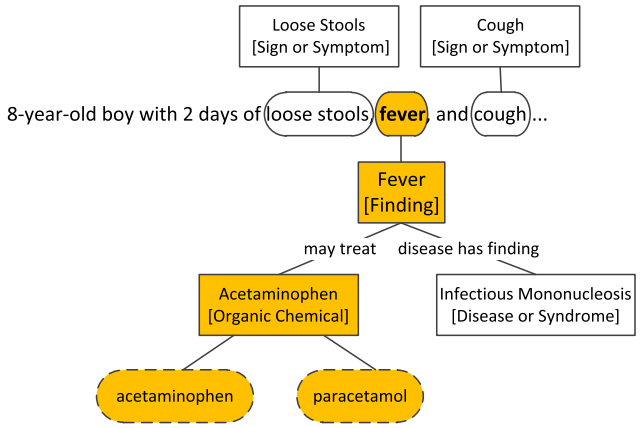


FIGURE 6. The example of knowledge-based query expansion.

is one of the top-ranked semantic associate triples obtained from the CP decomposition. This triple will be used as the expansion path, as indicated in Figure 6. New concepts like "Acetaminophen" are used as expansion concepts. Finally, the terms associated with this concept will be added to original query as expansion terms (terms in dotted circle in Figure 6). For each query, we set the maximum number of expansion terms based on the specific data.

V. EVALUATION

A. DATASET

The dataset we used to evaluate the performance of the proposed medical IR system is the 2014 TREC CDS Track [24], which aims at providing information relevant for patient care by linking the medical documents with the queries generated by electronic health records. The task is to answer generic clinical questions by retrieving relevant biomedical articles. The full dataset includes test queries, document collection, and relevance judgment file (available at http://trec.nist.gov/data/clinical2014.html). The document collection is a subset (733,138 articles) of the PubMed central, an online digital database of freely available biomedical literatures. The test queries are short medical case reports that represent actual medical records. Each case report is classified as one of the three types including diagnosis, treatment and test according to their different clinical information needs. A sample case report in XML format is shown in Table 2. For this case, "topic number" refers to its unique ID; "type" represents that the purpose of this case is to seek information for diagnosis; "description" provides the

TABLE 2. Example of TREC CDS 2014 queries.

Topic number	Medical Query Information
2	*<topic number="2" type="diagnosis"> <description>An 8-year-old male presents in March to the ER with fever up to 39 C, dyspnea and cough for 2 days. He has just returned from a 5 day vacation in Colorado. Parents report that prior to the onset of fever and cough, he had loose stools. He denies upper respiratory tract symptoms. On examination he is in respiratory distress and has bronchial respiratory sounds on the left. A chest x-ray shows bilateral lung infiltrates.</description> <summary>8-year-old boy with 2 days of loose stools, fever, and cough after returning from a trip to Colorado. Chest x-ray shows bilateral lung infiltrates.</summary> </topic>*
21	*<topic number="21" type="treatment"> <description> A 21-year-old female is evaluated for progressive arthralgias and malaise. On examination she is found to have alopecia, a rash mainly distributed on the bridge of her nose and her cheeks, a delicate non-palpable purpura on her calves, and swelling and tenderness of her wrists and ankles. Her lab shows normocytic anemia, thrombocytopenia, a 4/4 positive ANA and anti-dsDNA. Her urine is positive for protein and RBC casts. </description> <summary> 21-year-old female with progressive arthralgias, fatigue, and butterfly-shaped facial rash. Labs are significant for positive ANA and anti-double-stranded DNA, as well as proteinuria and RBC casts. </summary> </topic>*

detailed information of the case; whereas the ''summary'' is the simplified version which contains less irrelevant information. The task is to retrieve a set of medical documents to support the diagnosis of this case through identifying the relevance between the query (''summary'' or ''description'') and the documents from PubMed central. In this research, we evaluated the medical IR performance of all the 30 queries provided by this track, including 10 queries for each of the three types.

B. EXPERIMENT SETUP

The procedure of the experiment is as follows. In the preprocessing step, we extracted the information from the original XML files of queries and documents. Solr [23], a Lucene-based full-text search engine, is then used to build the index for all documents. The text retrieval algorithm BM25 was used for the initial ranking of documents. To build the proposed latent semantic relevance model introduced in Section IV.B, semantic association triples are extracted from queries and documents labeled as relevant or irrelevant. MetaMap [42] is used for concept mapping from the text in the queries and documents to UMLS knowledge base. Next, the system will retrieve the semantic relations between the concepts in the semantic network of UMLS knowledge base. The proposed rank estimation and CP decomposition algorithms are implemented based on the basic functions in a public Python library for multi-linear algebra and tensor factorizations [43].

For cross-validation, we divided test queries into three groups according to their types for diagnosis, test and treatment respectively. For each query used for training, all documents labeled as relevant are included in the training set, and the same number of irrelevant documents are selected randomly to be included in the training set. Only summary section of each medical case report is used as the initial input for retrieval. For parameter settings, incremental ratio for incremental PRF and number of expansion terms for knowledge-based query expansion are learned with the polling method using the training set.

C. EVALUATION RESULTS

The evaluation of the proposed medical IR system follows the standard TREC evaluation method for ad hoc retrieval tasks. Because it is not feasible to obtain complete relevance judgements for this huge collection of documents, inferred measures were adopted by TREC evaluating the performance. In particular, documents with a high relevance score (judged by competing systems) were sampled for human experts to judge the relevance as ''not relevant,'' ''possible relevant,'' or ''definitely relevant.'' These human labeled samples were then used to evaluate the performance of competing systems. In this study, we followed TREC standard and adopted the inferred Normalized Discounted Cumulative Gain (infNDCG) and inferred Average Precision (infAP) [44]. The performance for 3-fold cross-validation is shown in Figure 7. We evaluated the performance of three systems – the baseline system of Solr, the PRF system that incorporated the incremental pseudo relevance feedback introduced in Section IV.A, and the proposed system that incorporated both the PRF and knowledge-based query expansion system introduced in Section IV. The results demonstrated that the newly proposed methods demonstrated significant improvement over the baseline (an improvement of 38.65% to 100.81% across the 3 groups). The sample of identified semantic triples, top 1000 documents list for each query and evaluation results are available online (https://github.com/wanghaolin/SemanticMedIR/tree/master/data).

In addition to infNDCG and infAP, we also adopted the standard Precision at rank 10 (P10) provided by the official evaluation program of TREC. The results of all three evaluation methods are shown in Table 3. The proposed system with the two newly introduced methods achieved the best performance among the three systems. In Table 4, we compared the results of the proposed system with the performance of the top 3 results in terms of infNDCG reported by Roberts, Simpson *et al.* [25]. The results demonstrated that the performance of the proposed system outperformed state-of-the-art approaches on this retrieval task. However, it's worth noting, as recommended by TREC, cross-system comparisons should be avoided because different approaches have their unique novelty from different perspectives. The superior performance of the proposed system showed the potential of incorporating knowledge

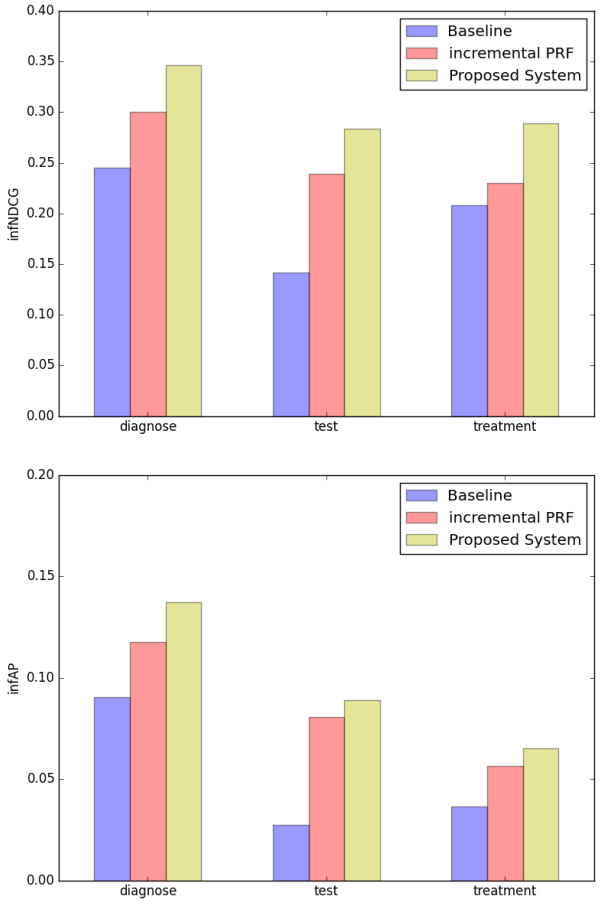


FIGURE 7. The performance of the baseline and proposed medical IR systems.

TABLE 3. Overall performances of proposed methods.

	infNDCG	infAP	P10
Solr (baseline)	0.1981	0.0514	0.2911
Incremental PRF	0.2565	0.0848	0.3600
The proposed system	**0.3062**	**0.0971**	**0.4100**

TABLE 4. Top 3 results of TREC CDS 2014 (by infNDCG).

	infNDCG	infAP	P10
1	0.2674	0.0659	0.3633
2	0.2631	0.0757	0.3900
3	0.2587	0.0812	0.3700

bases using tensor factorization to enhance medical IR methods.

As an active research area, a variety of approaches are explored recently with the same dataset to optimize the performance of medical information retrieval. Our experiment results are comparable with the state-of-the art systems developed by Zheng and Wan [45] and Balaneshin-kordan and Kotov [46]. In particular, the method presented in [46] and our method used both explicit and latent concepts from documents and knowledge bases. It is worth noting that the two approaches tackled the problem from two perspectives. The method in [46] focused on optimizing the query expansions using different sources to improve the retrieval performance. In our system, we used tensor factorization to extract meaningful semantic associations, rather than explicitly focusing on the retrieval performance. Both approaches are important to our understanding of how to use knowledge bases for information retrieval. It is an interesting future research topic to explore the feasibility to integrate both approaches to develop a more accurate and reliable medical information retrieval system.

VI. CONCLUSION

In this research, we proposed a medical IR system with a two-stage query expansion strategy, based on the incorporation of semantics from knowledge bases with tensor factorization methods. Experiments with the TREC dataset demonstrated the effectiveness of the proposed system. The proposed system has the potential to be adapted in other machine learning and medical informatics applications, like recommender systems, ontology learning, bioinformatics, etc. In our future research, we will (a) evaluate the performance of the proposed system with other medical IR datasets; (b) explore the feasibility of integrating the proposed tensor-based latent semantic relevance model with the probabilistic tensor decomposition framework to further enhance the performance of the medical IR system through introducing a Bayesian approach [32],[47]; (c) implement the system in real-world medical decision support applications through the collaboration with doctors and decision makers in local hospitals.

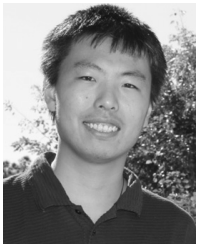
REFERENCES

[1] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," in *Biomedical Informatics*. New York, NY, USA: Springer, 2014, pp. 643–674.
[2] L. Goeuriot, L. Kelly, G. J. Jones, H. Müller, and J. Zobel, "Report on the SIGIR 2014 workshop on medical information retrieval (MedIR)," *ACM SIGIR Forum*, vol. 48, no. 2, pp. 78–82, 2014.
[3] G. Zuccon, B. Koopman, and P. Bruza, "Exploiting inference from semantic annotations for information retrieval: Reflections from medical IR," in *Proc. 7th Int. Workshop Exploiting Semantic Annotations Inf. Retr.*, 2014, pp. 43–45.
[4] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Comput. Surv.*, vol. 44, no. 1, p. 1, 2012.
[5] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 926–934.
[6] W. Shen and J.-Y. Nie, "Is concept mapping useful for biomedical information retrieval?" in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.*, 2015, pp. 281–286.
[7] B. Ermiş, E. Acar, and A. T. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Mining Knowl. Discovery*, vol. 29, pp. 203–236, 2015.
[8] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 1, pp. 6–20, Jan. 2009.
[9] D. Rafailidis and P. Daras, "The TFC model: Tensor factorization and tag clustering for item recommendation in social tagging systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 3, pp. 673–688, May 2013.
[10] D. Rafailidis and A. Nanopoulos, "Modeling users preference dynamics and side information in recommender systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 6, pp. 782–792, Jun. 2016.
[11] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 1, pp. 129–142, Jan. 2015.

[12] S. Rendle, "Factorization machines," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 995–1000.
[13] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, Jan. 2016.
[14] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, pp. D267–D270, Jan. 2004.
[15] Z. Liu and W. W. Chu, "Knowledge-based query expansion to support scenario-specific retrieval of medical free text," *Inf. Retr.*, vol. 10, no. 2, pp. 173–202, 2007.
[16] D. Martinez, A. Otegi, A. Soroa, and E. Agirre, "Improving search over electronic health records using UMLS-based query expansion through random walks," *J. Biomed. Inform.*, vol. 51, pp. 100–106, Oct. 2014.
[17] A. Otegi, X. Arregi, O. Ansa, and E. Agirre, "Using knowledge-based relatedness for information retrieval," *Knowl. Inf. Syst.*, vol. 44, no. 3, pp. 689–718, 2015.
[18] P. Sfakianaki *et al.*, "Semantic biomedical resource discovery: A natural language processing framework," *BMC Med. Inform. Decision Making*, vol. 15, p. 77, Sep. 2015.
[19] J. Mao, K. Lu, X. Mu, and G. Li, "Mining document, concept, and term associations for effective biomedical retrieval: Introducing MeSH-enhanced retrieval models," *Inf. Retr. J.*, vol. 18, no. 5, pp. 413–444, 2015.
[20] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley, "Information retrieval as semantic inference: A graph inference model applied to medical search," *Inf. Retr. J.*, vol. 19, no. 1, pp. 6–37, 2016.
[21] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action: Covers Apache Lucene 3.0*. Greenwich, CT, USA: Manning Publications Co., 2010.
[22] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft, "Indri: A language model-based search engine for complex queries," in *Proc. Int. Conf. Intell. Anal.*, 2005, pp. 2–6.
[23] T. Grainger, T. Potter, and Y. Seeley, *Solr in Action*. Greenwich, CT, USA: Manning, 2014.
[24] M. S. Simpson, E. M. Voorhees, and W. Hersh, "Overview of the TREC 2014 clinical decision support track," in *Proc. TREC*, Nov. 2014.
[25] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh, "State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS track," *Inf. Retr. J.*, vol. 19, no. 1, pp. 113–148, 2016.
[26] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.
[27] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," presented at the 22nd Int. Conf. Mach. Learn., Bonn, Germany, 2005.
[28] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Temporal link prediction using matrix and tensor factorizations," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 2, p. 10, 2011.
[29] E. Frolov and I. Oseledets. (2016). "Tensor methods and recommender systems." [Online]. Available: https://arxiv.org/abs/1603.06038
[30] T. Franz, A. Schultz, S. Sizov, and S. Staab, "TripleRank: Ranking semantic Web data by tensor decomposition," in *Proc. Int. Semantic Web Conf.*, 2009, pp. 213–228.
[31] K.-W. Chang, W.-T. Yih, B. Yang, and C. Meek, "Typed tensor decomposition of knowledge bases for relation extraction," in *Proc. EMNLP*, 2014, pp. 1568–1579.
[32] M. Nakatsuji, H. Toda, H. Sawada, J. G. Zheng, and J. A. Hendler, "Semantic sensitive tensor factorization," *Artif. Intell.*, vol. 230, pp. 224–245, Jan. 2016.
[33] J. B. Kruskal, "Rank, decomposition, and uniqueness for 3-way and N-way arrays," in *Multiway Data Analysis*, vol. 33. Dordrecht, The Netherlands: Elsevier, 1989, pp. 7–18.
[34] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 24–40, 2011.
[35] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2008, pp. 243–250.
[36] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," in *Text Mining*. Chichester, U.K.: Wiley, 2010, pp. 1–20.
[37] S. Al-Saffar, C. Joslyn, and A. Chappell, "Structure discovery in large semantic graphs using extant ontological scaling and descriptive semantics," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Aug. 2011, pp. 211–218.
[38] Q. Zhang and D. Haglin, "Semantic similarity between ontologies at different scales," *IEEE/CAA J. Autom. Sinica*, vol. 3, no. 2, pp. 132–140, Apr. 2016.
[39] P. Comon, X. Luciani, and A. L. F. de Almeida, "Tensor decompositions, alternating least squares and other tales," *J. Chemometrics*, vol. 23, pp. 393–405, Jul./Aug. 2009.
[40] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 809–816.
[41] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations with missing data," in *Proc. SDM*, 2010, pp. 701–712.
[42] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, May 2010.
[43] M. Nickel. (2013). *Scikit-Tensor Library*. [Online]. Available: http://pypi.python.org/pypi/scikit-tensor
[44] E. Yilmaz, E. Kanoulas, and J. A. Aslam, "A simple and efficient sampling method for estimating AP and NDCG," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2008, pp. 603–610.
[45] Z. Zheng and X. Wan, "Graph-based multi-modality learning for clinical decision support," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 1945–1948.
[46] S. Balaneshin-Kordan and A. Kotov, "Optimization method for weighting explicit and latent concepts in clinical decision support queries," in *Proc. ACM Int. Conf. Theory Inf. Retr.*, 2016, pp. 241–250.
[47] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with Bayesian probabilistic tensor factorization," in *Proc. SIAM Int. Conf. Data Mining*, 2010, pp. 211–222.

HAOLIN WANG received the B.S. degree in electronic and information engineering from Beihang University and the M.S. degree in software engineering from the University of Science and Technology of China. He is currently pursuing the joint Ph.D. degree with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, and with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences, Beijing, China.

His research interests include data mining, machine learning, and information retrieval.



QINGPENG ZHANG (M'10) received the B.S. degree in automation from the Huazhong University of Science and Technology, and the M.S. degree in industrial engineering and the Ph.D. degree in systems and industrial engineering from The University of Arizona. He was a Post-Doctoral Research Associate with The Tetherless World Constellation, Rensselaer Polytechnic Institute. He is currently an Assistant Professor with the Department of Systems Engineering and Engineering Management, City University of Hong Kong. He is an Associate Editor of the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

His research interests include social computing, complex networks, healthcare data analytics, semantic social networks, and web science.

JIAHU YUAN received the B.S. degree from the Huazhong University of Science and Technology, the M.S. degree from the Chinese Academy of Sciences, and the Ph.D. degree from Sichuan University. He is currently the Director and a Professor with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences.

His research interests include artificial intelligence and intelligent systems.