CTGA: Graph-based Biomedical Literature Search

Tianwen Jiang[†,‡*], Zhihan Zhang[†,§*], Tong Zhao[†], Bing Qin[‡], Ting Liu[‡], Nitesh V. Chawla[†], Meng Jiang[†]
[†]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
[‡]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
[§]School of Electronic Engineering and Computer Science, Peking University, China
twjiang@ir.hit.edu.cn, zhangzhihan@pku.edu.cn, tzhao2@nd.edu, {bqin, tliu}@ir.hit.edu.cn, {nchawla, mjiang2}@nd.edu

***Abstract*—Scientists are relying heavily on biomedical literature search (BLS) engines (e.g., PubMed) to acquire knowledge. Existing BLS systems adopt a "C-A" paradigm that is to design query-document similarity measurement based on words/phrases in the unstructured Content and to develop search Algorithms. In this work, we argue that structures should be extracted and utilized to bridge the gap between text content and knowledge-based search. And *graph* is one of the most effective *structured* forms of knowledge and more informative than words or phrases. So we carry out a paradigm shift from "C-A" to "CTGA". Here "T" is for factual tuple of concepts and relations, and "G" is for knowledge graph. Our proposed graph-based BLS system has three parts: (1) it uses neural information extraction models to turn text into tuples; (2) it represents the tuples of a query or a document as a knowledge graph of linked concept nodes; (3) it has an efficient graph-based matching algorithm to return related documents at the level of structured knowledge. Experiments show that in both objective and subjective evaluation, our CTGA performs significantly better than traditional CA-based PubMed.**

I. Introduction

Biomedical literature search (BLS) systems are essential for providing bio-scientists with quick access to large biomedical literature database and effective search for useful information. As biomedical research develops rapidly in this era, it becomes a growing challenge for scientists to stay up to date with the latest advances in the domain. Those BLS systems are of great help in guiding researchers to formulate or validate hypotheses and discover knowledge.

Various approaches have been made towards the development of BLS system. The most widely used service is PubMed provided by National Center for Biotechnology Information (NCBI). Other approaches include GoPubMed [1], HubMed [2], ReleMed [3] and iPubMed [4], which make some improvements on the efficiency and precision of PubMed. More recent approaches [5]–[7] introduce natural language processing (NLP) tools into search systems.

Unfortunately, existing BLS search engines have limitations to return knowledge for users as needed. They designed query-document similarity measurements based on words or phrases in the unstructured Content and developed search Algorithms. Such a "C-A" paradigm is not satisfactory, because keywords or concepts are too limited to represent rich structures of knowledge in biomedical statements. The retrieved articles with matched words may fail to provide relevant information to the input query. For instance, given the query as "... *VPA treatment increased cell proliferation* ...", the user might expect retrieved documents to describe what had effects on cell proliferation, increasing or even reducing, and/or to describe what else observed effects VPA treatment could generate. However, the word-based search systems can only return documents with the word *VPA* or *cell* or the phrase *cell proliferation* appearing many times, which may deviate from the expectation.

In this work, we argue that structured representations of knowledge should be extracted and utilized to bridge the gap between text content and knowledge-based BLS. In this paper, we suggest to use Tuples and Graphs as the structured representations of biomedical knowledge. Take the query example again. If the query could be transformed into a tuple as *(VPA treatment, increased, cell proliferation)* and if a great number of tuples related to the same or relevant concepts could be linked together as a graph, we would have a more structured and informative form of knowledge to find relevant documents to the query. Specifically, on one side, information extraction (IE) techniques have defined (subject, relation, object)-tuple as a standard format of the task outcome. Associations between concepts and relations can naturally be represented as linked graph data. On the other side, a wide line of graph representation learning and graph matching algorithms have been proposed and widely applied in various data such as social networks and biological networks [8]–[10]. When transforming text into tuples and graphs, those algorithms become great tools for search, learning, and exploration.

Therefore, we carry out a paradigm shift from the traditional word-based search towards graph-based search. We propose "CTGA", a novel **C**ontent → **T**uple → **G**raph → **A**lgorithm paradigm for BLS. First, we apply MIMO, a deep neural information extraction (IE) model [11] to extract factual and conditional tuples from unstructured content. Second, we construct a three-layer knowledge graph based on the extracted information as the representation of documents in our database. Third, we design a graph-based search algorithm to compare the knowledge graphs between the query and the documents and to return the most relevant documents. Experiment results from both objective and subjective evaluation suggest that our proposed CTGA outperforms the PubMed system. Further case study demonstrates that CTGA can assist researchers to find relevant information and discover knowledge. The system is now launched at http://www.biokgs.com.

The main contributions can be summarized as follows:

- We propose CTGA, a graph-based biomedical literature

*Equal contribution.

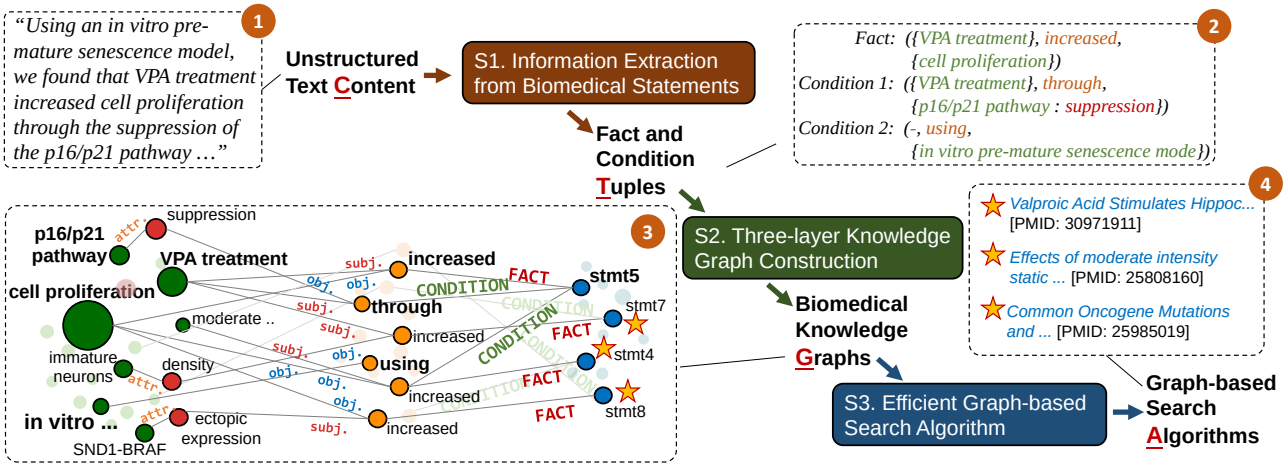

Fig. 1: The workflow of CTGA. There are three steps for graph-based BLS: *S1* extracts fact and condition **T**uples (see ②) from unstructured text **C**ontent (see ①); *S2* constructs three-layer knowledge **G**raphs (see ③) using the tuples; S3 performs a graph-based search **A**lgorithm to retrieve the most relevant documents to query (see ④).

search engine with the novel "**C**ontent, **T**uple, **G**raph, and **A**lgorithm" paradigm.

- We design a novel form, three-layer graph, to represent knowledge in literature as well as an efficient graph-based matching algorithm to perform knowledge-level search.
- Experimental results show that our CTGA outperforms PubMed in both objective and subjective evaluation, leading bio-scientists to discover knowledge in literature.

II. Proposed Method

Our goal is to perform biomedical literature search on structured form of content. Graphs are more structured and informative than keywords. And the knowledge graphs can be constructed from the text content in an information-lossless way, if concepts and relations are extracted. So our BLS system is graph-based and has a novel paradigm called **CTGA**. It is in short of **C**ontent, **T**uple, **G**raph, and **A**lgorithm. It is a fun fact that the initials are the same as the nitrogenous bases in DNA molecules. Fig. 1 illustrates the workflow of CTGA.

A. S1: From Content to Tuples

Here we describe how to extract structured information from unstructured text. We use *(subject, relation, object)*-tuples as the basic structure. The subject or object is composed of a concept and its attribute (if exists). The relation is often a predicate. Formally, We denote the tuples as

$$t = (\{c_1 : a_1\}, r, \{c_3 : a_3\}), \tag{1}$$

where c_1 and a_1 are the concept and attribute of the subject, c_3 and a_3 are the concept and attribute of the object, r stands for the relation between the subject and the object.

Traditional IE tasks focus on extracting factual information only. However, biomedical literature may have a number of clauses that serve as important conditional information of the facts. Thus, we extract both *fact tuples* and *condition tuples* from statement sentences. We adopt the state-of-the-art neural biomedical IE model, called multi-input multi-output sequence labeling (MIMO) [11]. It has an encoder-decoder framework with multi-input and multi-output modules. MIMO takes a word sequence as input and performs multi-class classification to predict a tag for each word. The tuple-oriented tag schema [11] allows us to find fact tuples and condition tuples.

B. S2: From Tuples to Knowledge Graph

Tuples are structured but isolated. Graph can link things such as concepts, attributes, and relations in the tuples together. So, in this section we present a set of notations and definitions to turn tuples into knowledge graph. An example of the knowledge graph can be found at the bottom left corner of Fig. 1. First, we define the graph form of a particular tuple $t = (\{c_1 : a_1\}, r, \{c_3 : a_3\})$ along with its source statement s:

$$G_t = < V_t, E_t > \tag{2}$$
$$V_t = \{v_t^{c_1}, v_t^{a_1}, v_t^r, v_t^{c_3}, v_t^{a_3}, v_t^s\} \tag{3}$$
$$E_t = \{e_t^{subj}, e_t^{obj}, e_t^{stmt}, e_t^{attr_1}, e_t^{attr_3}\} \tag{4}$$

where V_t and E_t represent the set of nodes and the set of edges in G_t. v_t^x represents the corresponding node of tuple unit x ($x \in \{c_1, c_3, r, a_1, a_3\}$). v_t^s denotes the source statement from which the tuple is extracted. e_t^{stmt} denotes for the edge between v_t^s and v_t^r. $e_t^{attr_x}$ denotes the edge between $v_t^{c_x}$ and $v_t^{a_x}$ ($x = \{1, 3\}$). $e_t^{subj} = (v_t^r, v_t^{c_1})$ is the edge between the subject $v_t^{c_1}$ and the relation v_t^r. $e_t^{obj} = (v_t^r, v_t^{c_3})$ is the edge connecting the relation v_t^r with the object $v_t^{c_3}$.

Subsequently, the knowledge graph G_d of a given document d can be represented as the union of G_t:

$$G_d = \bigcup_{t \in T_d} G_t \tag{5}$$

where T_d is the tuple set extracted from document d. Finally, we construct a global knowledge graph which can be repre-

sented as the union of each document's graph:

$$BioKG = \bigcup_{d \in DB} G_d \tag{6}$$

where DB denotes the database of all documents.

We use $V_{\mathcal{C}}, V_{\mathcal{A}}$, $V_{\mathcal{R}}$, and $V_{\mathcal{S}}$ to denote the nodes of concepts, attributes, relations, and statement sentences, respectively. Edge sets $E_{\mathcal{CA}}$, $E_{\mathcal{RC}}$, and $E_{\mathcal{SR}}$ are used in short of $\{e_t^{attr_x} \mid t \in T, x = 1, 3\}$, $\{e_t^y \mid t \in T, y = subj, obj\}$, and $\{e_t^{stmt} \mid t \in T\}$, respectively, where T is the set of tuples in all documents. Formally, the proposed knowledge graph is defined in form of node sets and edge sets:

$$BioKG =< V_{\mathcal{C}} \cup V_{\mathcal{A}} \cup V_{\mathcal{R}} \cup V_{\mathcal{S}},\ \ E_{\mathcal{CA}} \cup E_{\mathcal{RC}} \cup E_{\mathcal{SR}} > . \tag{7}$$

As illustrated in Fig. 1, our knowledge graph is organized in a three-layer structure: the first layer consists of concept and attribute nodes $V_{\mathcal{C}}$, $V_{\mathcal{A}}$; the second layer consists of relation nodes $V_{\mathcal{R}}$; the third layer consists of statement nodes $V_{\mathcal{S}}$. Layer 1 and 2 are linked by edges in $E_{\mathcal{PC}}$, indicating the association between relations and subjects/objects. Layer 2 and 3 are linked by edges in $E_{\mathcal{SP}}$, indicating the source statement of each fact or condition tuple. The advantage of such layout is that tuples in a certain statement can be easily detected through their relation node, making it easy to access the conditions from a fact on the knowledge graph.

C. S3: From Graph to Search Algorithms

In this section, we describe a graph-based search algorithm to retrieve the most relevant documents from the database and rank them by their relevance to the query. We first consider a certain *concept-relation-concept* form of path, which is defined as an ordered sequence of nodes and edges in the graph:

$$\mathcal{P} = (v^{a_1}, e^{attr_1}, v^{c_1}, e^{subj}, v^r, e^{obj}, v^{c_3}, e^{attr_3}, v^{a_3}). \tag{8}$$

We name such paths $\mathcal{P}$ in the knowledge graph as ***KC paths***, where KC is for "**K**nowledge **C**arrier". KC paths in G_q and G_d will be used to calculate the relevance between the graphs.

We use modified *Precision* and *Recall* scores to calculate the relevance between the query's graph G_q and the document's graph G_d. A high *Precision* means that KC paths in the document are *highly similar* to those in the query, while a high *Recall* shows that KC paths in the query are *completely covered* by those in the document. The general relevance between the query's graph and the document's graph is defined as the harmonic mean of *Precision* and *Recall*:

$$\text{Precision}_{G_q,G_d} = \frac{\sum_{i=1}^{N} \max_{j=1,2,\ldots,M} \phi(\mathcal{P}_i, \mathcal{P}_j)}{N} \tag{9}$$

$$\text{Recall}_{G_q,G_d} = \frac{\sum_{j=1}^{M} \max_{i=1,2,\ldots,N} \phi(\mathcal{P}_i, \mathcal{P}_j)}{M} \tag{10}$$

$$\text{Relevance}_{G_q,G_d} = (1+\beta^2)\frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{11}$$

where N and M denote the number of detected KC paths in the query and the document respectively. $\phi(\mathcal{P}_i, \mathcal{P}_j)$ is the function for computing the similarity of a pair of KC paths. β is a hyper-parameter for balancing the importance of *Precision* and *Recall*.

As the implementation of $\phi(\mathcal{P}_i, \mathcal{P}_j)$, we calculate the similarity between a pair of KC paths $\mathcal{P}_i$ and $\mathcal{P}_j$ by summing the similarity score of each pair of nodes:

$$\begin{aligned}\phi(\mathcal{P}_i, \mathcal{P}_j) = &\text{sim}^c(v_i^{c_1}, v_j^{c_1}) + \text{sim}^w(v_i^{a_1}, v_j^{a_1}) \\ &+ \text{sim}^c(v_i^{c_3}, v_j^{c_3}) + \text{sim}^w(v_i^{a_3}, v_j^{a_3}) \\ &+ \text{sim}^w(v_i^r, v_j^r)\end{aligned} \tag{12}$$

where sim^c or sim^w is the similarity function of a pair of nodes. Particularly, sim^c computes concept similarity:

$$\text{sim}^c(v_i^x, v_j^x) = \begin{cases} 1, & if\ \text{Synonym}(x^i, x^j) \\ \mathbf{x}^i \cdot \mathbf{x}^j, & if\ \text{ShareHypernym}(x^i, x^j) \\ 0, & otherwise \end{cases} \tag{13}$$

where $\mathbf{x}^i$ and $\mathbf{x}^j$ are the embedding representations of the corresponding text x^i and x^j, $x \in \{c_1, c_3\}$. "Synonym" or "ShareHypernym" means whether x^i and x^j are synonyms or share the same hypernym with each other, which can be determined through the search system of the popular ontology database Ontobee.

Similarly, sim^w computes similarity for attribute and relation nodes:

$$\text{sim}^w(v_i^x, v_j^x) = \mathbf{x}^i \cdot \mathbf{x}^j \tag{14}$$

where $\mathbf{x}^i$ and $\mathbf{x}^j$ are the embedding representations of the corresponding text x^i and x^j, $x \in \{a_1, a_3, r\}$. The difference between sim^w and sim^c is that concepts are scientific entities which contain more semantic information than common words, thus we use external ontology database to assist computing similarity between concepts.

III. Experiments

A. Baseline System: PubMed

We conduct experiments to compare our proposed CTGA and the CA-based system PubMed, the most widely-used public resource for BLS. Users are supposed to compose a keyword query and PubMed returns a list of the most relevant documents from its database. The documents are ranked in order, with most relevant ones on the top.

B. Data Processing

We collected 15.5 million scientific articles in biomedical domain from MEDLINE as the CTGA's database. Documents are first processed by the MIMO model to extract structured information and then transformed into knowledge graphs. For search efficiency, the knowledge graphs are managed in MySQL databases. Indices for KC paths are created across records to equip CTGA with quick and robust performance.

C. Experiment Settings

We use the documents listed as "Latest literature" and "Trending articles" on PubMed to collect our test data. We ask 5 domain experts to select biomedical statements they are interested in, making up a total number of 100 statements

TABLE I: Experiment results of objective evaluation. Suffix "**-FullText**" means using full text of the retrieved documents in evaluation and suffix "**-BestMatch**" means using the best-match sentence in each document. **SkipThought** and **Aver.Embedding** represent average cosine similarity of Skip-Thought vectors and GloVe embedding vectors, respectively.

Search Engines	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	SkipThought	Aver.Embedding	Top@5	Top@1	MRR
PubMed-FullText	17.22	7.01	4.43	3.37	16.34	75.56	68.54	88%	77%	81.5
CTGA-FullText	**17.61**	**7.73**	**5.07**	**3.90**	**17.03**	**75.91**	**68.77**	**93%**	**92%**	**92.3**
Improvement	2.3%↑	10.3%↑	14.4%↑	15.7%↑	4.2%↑	0.5%↑	0.3%↑	5.7%↑	19.5%↑	13.3%↑
PubMed-BestMatch	28.84	21.46	18.42	16.71	34.91	80.31	81.92	-	-	-
CTGA-BestMatch	**32.60**	**24.67**	**21.06**	**18.84**	**38.26**	**81.77**	**82.80**	-	-	-
Improvement	13.0%↑	15.0%↑	14.3%↑	12.7%↑	9.6%↑	1.8%↑	1.1%↑	-	-	-

TABLE II: Experiment results of subjective evaluation. **Rate@**k means the average ratings of top k documents in the returned list. **Preference%** means the number of cases that each search engine wins in the two-side comparison. Each statement is reviewed by 5 annotators and the system with more votes is the winner of the single case. Lower **SF-Dist** performs better.

Search Engines	Rate@1	Rate@3	Rate@5	Preference%	Kendall-Tau	SF-Dist	RBO	DCG	NDCG
PubMed	4.60	3.43	2.92	34%	68.95	10.14	87.18	9.61	96.63
CTGA	**4.83**	**3.69**	**3.31**	**66%**	**76.15**	**8.34**	**89.31**	**10.60**	**99.25**

for testing. Note that these statements are collected from real publications so that each statement has its corresponding source article in the database. Subsequently, we use these 100 statements as queries to retrieve relevant documents in CTGA and PubMed respectively. We take the top 5 documents with their titles and abstracts as the search results.

Retrieved documents may contain lots of irrelevant information to the query. So, in each document we mark out a "best-match sentence" that matches the most number of content words (nouns, verbs, and adjectives) to the query. We use these sentences to assist for evaluating the performance.

D. Evaluation Metrics

We conduct experiments on both objective (data-based) evaluation and subjective (human-based) evaluation.

1) ***Objective Evaluation****:*

BLEU [12] and ROUGE [13] are reference-based metrics to measure the correspondence between the query and sentences in the retrieved documents. We use Skip-Thought vectors [14] and GloVe embeddings [15] to obtain the sentence representations of the query and documents, then compute their cosine similarity. We report the results between the query and the best-match sentence in each document. The intuition is that full-text scores show the overall relevance between the documents and the query, while best-match scores indicate how exactly the documents contain the query's information.

Since each query has its source article in the database, we measure the ranking quality of search engines by judging whether the source article is ranked on the top. We report Top@k, the percentage that the source article appears in the top k results in the retrieved document list, and MRR, the mean reciprocal rank of the source articles.

2) ***Subjective Evaluation****:*

We recruited 5 human annotators with biomedical expertise. Each statement will be reviewed for 5 times to reduce bias. In detail, the annotators rated each document on a scale of 1 to 5 (higher is better) based on information relevance between the document and the factual query. The annotators voted for his/her preferred side from two sets of search results. The document ratings represent the individual relevance of each retrieved document, and the annotators' preferences show the overall quality of the search results. We also compare the similarity between the document rankings given by the two systems with the rankings given by annotators (sorted by ratings in descending order). We calculate Kendall Tau coefficient [16], Spearman's Footrule distance (SF-Dist) [16], and rank-biased overlap (RBO) [17] to measure the ranking effectiveness. Finally, Discounted Cumulative Gain (DCG) [18] and Normalized Discounted Cumulative Gain (NDCG) [18] are calculated to test both the ranking quality and the document relevance by the two systems.

E. Experimental Results

Table I and II present experimental results of objective and subjective evaluation, respectively. The results indicate that the proposed CTGA outperforms PubMed in all the evaluation metrics. We further analyze the results in the following aspects.

a) ***Relevance of retrieved documents****:* BLEU, ROUGE and similarity scores in Table I and Rate@k scores in Table II indicate higher relevance between the queries and retrieved documents from CTGA than those from PubMed. Interestingly, the results from BLEU and ROUGE show much larger margin using the best-match sentences than using the full text. It demonstrates that CTGA performs better at accurately retrieving relevant knowledge. This is because CTGA conducts graph-based search rather than simply matching as many keywords as possible. Our graph-based paradigm enables CTGA to retrieve articles with sentences containing closely related information to the query.

b) ***Effectiveness of article ranking****:* Table I shows that CTGA performs better on assigning higher ranks to source articles in the returned list according to Top@k scores and MRR. Since the source article contains the exact match of

the query, CTGA shows stronger capabilities in discriminating relevant documents from irrelevant ones. It indicates CTGA has a better ranking ability. Besides, Kendall Tau coefficient, Spearman's Footrule distance, and RBO scores in Table II suggest that the ranking orders of CTGA have closer relevance with human needs than those of PubMed. This demonstrates the ranking effectiveness of CTGA.

*c) **Overall quality of search engines**:* Reviewer preference percentage, DCG and NDCG scores in Table II measure the overall quality of the two search engines. In 66% of the cases, the search results of CTGA receive more favor from human annotators, while PubMed has only 34% winning cases. Furthermore, CTGA wins by 5-0 in 35 cases out of 66, while the ratio of PubMed is only 7 out of 34. This shows that human annotators often reach consensus when CTGA outperforms PubMed, while they cannot reach an unanimous agreement most of the time when PubMed wins the votes. As for inter-annotator agreement, the average Pearson correlation coefficient of human ratings is 0.71 (ranging from -1 to 1, higher is better), suggesting that the rating distributions are highly consistent. Besides, CTGA outscores PubMed in DCG as well as NDCG, which indicates more effective search algorithms and better document quality. In general, CTGA achieves better results in various evaluation metrics because of its novel graph-based search algorithm and robust ranking ability.

F. Case Study

Fig. 2 presents a case study of the search results from CTGA and PubMed. The user inputs the query "VPA treatment increased cell proliferation" and hopes to obtain relevant knowledge about this fact. CTGA performs graph-based search on knowledge level so that each retrieved document contains certain knowledge correlated with the query. On the other hand, PubMed conducts word-based search so the retrieved documents all contain several keywords from the query. However, PubMed cannot guarantee the quality and quantity of knowledge content in the retrieved documents so many documents may have overlap in keywords but express irrelevant facts to the query. Apparently, the results from CTGA are more likely to meet the user's need.

Additionally, we also discover great usefulness from the search results of CTGA. Besides retrieving information that is similar to the query "VPA treatment increased cell proliferation", CTGA also returns multiple documents containing interesting knowledge that the user would probably like to know. For instance, in addition to the fact that VPA treatment is able to increase cell proliferation, we can learn that VPA treatment can also inhibit apoptosis, increase the density of immature neurons or increase cathepsin B levels (Doc #1∼#3). We can also find other events or substances that will lead to an increase in cell proliferation, such as moderate intensity SMFs and ectopic expression of SND1-BRAF (Doc #4∼#5). All these obtained knowledge should be attributed to the graph-based search paradigm in CTGA, which can better assists bioscientists in their research works.

Query

VPA treatment increased cell proliferation.

CTGA

1. Using an in vitro pre-mature senescence model , we found that VPA treatment increased cell proliferation and inhibited apoptosis through the suppression of the p16/p21 pathway.
   (Knowledge: {VPA treatment, increase, cell proliferation}, {VPA treatment, inhibit, apoptosis})
2. VPA treatment promoted cell proliferation and increased the density of immature neurons in the dentate gyrus (DG) of the hippocampus of 3xTgAD mice.
   (Knowledge: {VPA treatment, promote, cell proliferation}, {VPA treatment, increase, density of immature neurons})
3. VPA treatment increased cathepsin B levels and activities in primary CLL cells.
   (Knowledge: {VPA treatment, increase, cathepsin B level})
4. Moderate intensity SMFs increased cell proliferation, ALP activity, and calcium release.
   (Knowledge: {Moderate intensity SMF, increase, cell proliferation})
5. Ectopic expression of SND1-BRAF in H1299 cells increased phosphorylation levels of MEK/ERK , cell proliferation , and spheroid formation
   (Knowledge: {Ectopic expression of SND1-BRAF, increase, cell proliferation})

PubMed

1. These results suggest that VPA increased type-1 stem cells in relation to the activation of SCF-KIT signaling and suppression of BTG2-mediated antiproliferative effect on stem cells.
2. Prostate cancer cells , sensitive and resistant to temsirolimus , were exposed to VPA , and tumor cell growth behavior compared.
3. VPA treatment promoted cell proliferation and increased the density of immature neurons in the dentate gyrus (DG) of the hippocampus of 3xTgAD mice.
4. Cell proliferation had increased to control levels at 30 and 45 d, demonstrating that memory recovery occurs over a period of six weeks after discontinuing VPA treatment.
5. To compare the protective effects of suberoylanilide hydroxamic acid (SAHA) and valproic acid (VPA) on human lens epithelial cells (HLECs) following ultraviolet-B exposure.

Fig. 2: Case study of search results from CTGA and PubMed. Words in red indicates matched keywords that appear in the query. Tuples marked as *"Knowledge"* in parentheses indicate scientific knowledge that can be inferred from CTGA results. Out of simplicity, we only present the best-match sentence of each document which contains the most relevant information to the query.

IV. Related Work

*a) **Biomedical search engines**:* Many search engines have achieved great success in biomedical field. The most well-known system is PubMed. FACTA [19] and Essie [20] selected relevant documents by searching for associated concepts with the query. ReleMed [3] further enhanced search precision by measuring inner-sentence relationships between words. iPubMed [4] developed novel index structures to perform interactive and fuzzy search on PubMed. BEST [6] extracted biomedical entities from the query and calculates document relevance based on entity co-occurence. LivTar [7] applied name recognition and entity recognition to conduct semantic-level searches. However, search algorithms based on either keywords or entities in these systems could not losslessly convey the knowledge in biomedical statements.

*b) **Biomedical information extraction**:* Information extraction is an important research topic in natural language processing [21]–[24]. Li *et al.* [25] used word embeddings as token features to perform bio-event extraction on biomedical text. De *et al.* [26] introduced semi-supervised learning into clinical information extraction. Soldaini *et al.* [27] improved clinical notes representation for document retrieval. Zheng *et al.* [28] applied multi-modal learning to retrieve biomedical literature for clinical decision support. Compared to previous work, our CTGA makes a breakthrough to apply a novel IE model to extract both factual and conditional information in biomedical literature.

*c) **Biomedical knowledge graphs**:* Integrated bio-entity network [29] was a graph data structure containing information of bio-entity relationships. BioGrakn [30], a graph-based deductive database, combined knowledge graph construction with machine reasoning. SemaTyP [31] constructed a knowledge graph from biomedical documents and applied it on drug discovery. Most existing knowledge graphs used concept-relation structures. However, our CTGA splits relations, concepts, attributes, and statements into different layers. Not only can this structure contain richer information, but it is also more effective for visualization.

V. Conclusions

In this work, we proposed CTGA, a graph-based biomedical literature search engine. We carried out a paradigm shift from word-based search to graph-based search. We applied the state-of-the-art neural information extraction model to extract structured knowledge from text content. We designed a novel three-layer knowledge graph to represent knowledge in an information-lossless way. We further devised a graph-based search algorithm to retrieve relevant documents. Experimental results showed that our CTGA outperformed PubMed in both objective and subjective evaluation.

Acknowledgments

This work was supported in part by National Key Research and Development Program of China No. 2018YFB1005103, National Natural Science Foundation of China (NSFC) No. 61632011, and NSF Grant CCF-1901059.

References

[1] A. Doms and M. Schroeder, "Gopubmed: exploring pubmed with the gene ontology," *Nucleic acids research*, vol. 33, no. suppl_2, pp. W783–W786, 2005.

[2] M. A. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. A. Wooldridge, and J. Ye, "Biotext search engine: beyond abstract search," *Bioinformatics*, vol. 23, no. 16, pp. 2196–2197, 2007.

[3] M. S. Siadaty, J. Shu, and W. A. Knaus, "Relemed: sentence-level search engine with relevance score for the medline database of biomedical articles," *BMC medical informatics and decision making*, vol. 7, no. 1, p. 1, 2007.

[4] J. Wang, I. Cetindil, S. Ji, C. Li, X. Xie, G. Li, and J. Feng, "Interactive and fuzzy search: a dynamic way to explore medline," *Bioinformatics*, vol. 26, no. 18, pp. 2321–2327, 2010.

[5] H. Poon, C. Quirk, C. DeZiel, and D. Heckerman, "Literome: Pubmed-scale genomic knowledge base in the cloud," *Bioinformatics*, vol. 30, no. 19, pp. 2840–2842, 2014.

[6] S. Lee, D. Kim, K. Lee, J. Choi, S. Kim, M. Jeon, S. Lim, D. Choi, S. Kim, A.-C. Tan *et al.*, "Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature," *PloS one*, vol. 11, no. 10, p. e0164680, 2016.

[7] A. Allot, Y. Peng, C.-H. Wei, K. Lee, L. Phan, and Z. Lu, "Litvar: a semantic search engine for linking genomic variant data in pubmed and pmc," *Nucleic acids research*, vol. 46, no. W1, pp. W530–W536, 2018.

[8] S. Gold and A. Rangarajan, "A graduated assignment algorithm for graph matching," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 4, pp. 377–388, 1996.

[9] A. Dutta, J. Lladós, H. Bunke, and U. Pal, "Product graph-based higher order contextual similarities for inexact subgraph matching," *Pattern Recognition*, vol. 76, pp. 596–611, 2018.

[10] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Neural Information Processing Systems*, 2018, pp. 4800–4810.

[11] T. Jiang, T. Zhao, B. Qin, T. Liu, N. V. Chawla, and M. Jiang, "The role of "condition": a novel scientific knowledge graph representation and construction model," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019.

[12] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Association for Computational Linguistics*, 2002, pp. 311–318.

[13] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[14] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems*, 2015, pp. 3294–3302.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of Conference of Empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[16] R. Kumar and S. Vassilvitskii, "Generalized distances between rankings," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 571–580.

[17] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 4, p. 20, 2010.

[18] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[19] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Facta: a text search engine for finding associated biomedical concepts," *Bioinformatics*, vol. 24, no. 21, pp. 2559–2560, 2008.

[20] N. C. Ide, R. F. Loane, and D. Demner-Fushman, "Essie: a concept-based search engine for structured biomedical text," *Journal of the American Medical Informatics Association*, vol. 14, no. 3, pp. 253–263, 2007.

[21] W. Yu, Z. Li, Q. Zeng, and M. Jiang, "Tablepedia: Automating pdf table reading in an experimental evidence exploration and analytic system," in *The World Wide Web Conference*. ACM, 2019, pp. 3615–3619.

[22] X. Wang, H. Zhang, Q. Li, Y. Shi, and M. Jiang, "A novel unsupervised approach for precise temporal slot filling from incomplete and noisy temporal contexts," in *The World Wide Web Conference*. ACM, 2019, pp. 3328–3334.

[23] Q. Zeng, M. Yu, W. Yu, J. Xiong, Y. Shi, and M. Jiang, "Faceted hierarchy: A new graph type to organize scientific concepts and a construction method," in *Proceedings of Workshop on Graph-Based Natural Language Processing (TextGraphs)*, 2019.

[24] T. Jiang, T. Zhao, B. Qin, T. Liu, N. V. Chawla, and M. Jiang, "Multi-input multi-output sequence labeling for joint extraction of factand condition tuples from scientific text," in *Proceedings of Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2019.

[25] C. Li, R. Song, M. Liakata, A. Vlachos, S. Seneff, and X. Zhang, "Using word embedding for bio-event extraction," in *Proceedings of BioNLP 15*, 2015, pp. 121–126.

[26] B. De Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu, "Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 557–562, 2011.

[27] L. Soldaini, A. Yates, and N. Goharian, "Denoising clinical notes for medical literature retrieval with convolutional neural model," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 2307–2310.

[28] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, "Joint extraction of entities and relations based on a novel tagging scheme," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, 2017, pp. 1227–1236.

[29] L. Bell, R. Chowdhary, J. S. Liu, X. Niu, and J. Zhang, "Integrated bio-entity network: a system for biological knowledge discovery," *PloS one*, vol. 6, no. 6, p. e21474, 2011.

[30] A. Messina, H. Pribadi, J. Stichbury, M. Bucci, S. Klarman, and A. Urso, "Biograkn: a knowledge graph-based semantic database for biomedical sciences," in *Conference on Complex, Intelligent, and Software Intensive Systems*. Springer, 2017, pp. 299–309.

[31] S. Sang, Z. Yang, L. Wang, X. Liu, H. Lin, and J. Wang, "Sematyp: a knowledge graph based literature mining method for drug discovery," *BMC bioinformatics*, vol. 19, no. 1, p. 193, 2018.