Visualization of PICO Elements for Information Needs Clarification and Query Refinement

Wan-Tze Vong and Patrick Hang Hui Then

Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia
{wvong,pthen}@swinburne.edu.my

Abstract. The UMLS semantic types and natural language processing techniques were collectively utilized to extract PICO elements from the titles and abstracts of 114 MEDLINE articles. 24 sets of PICO elements were generated from the articles based on the derivation of, and the tokenization methods and weighting schemes applied to the elements. The similarity of the I and C elements (called jointly the "Interventions") between pairs of documents was calculated using 42 similarity/distance measures. Similar interventions were grouped together using complete-/average-/ward-link hierarchical clustering. The similarity measure, Yule, performed significantly better than other measures in identifying paired interventions derived from the titles and which had been pre-processed into single term and weighted by binary term-occurrence. The clustering algorithm, complete-link, provides the most appropriate structure for the visualization of interventions. Similarity-based clustering gave a higher mean average precision than random-baseline clustering (MAP = 0.4298 vs. 0.2364) over the 25 queries evaluated.

Keywords: Hierarchical Clustering, PICO Element, Query Refinement, Similarity Measure, Distance Measure.

1 Introduction

A focused, well-defined question warrants a high quality answer. The quality of answers returned by a question-answering (QA) system depends on the quality of questions posed by users. Doctors have difficulty in generating high quality questions that unambiguously and comprehensively defined their information needs [1]. The use of PICO (an acronym for Problem/Population, Intervention, Comparison and Outcome) framework has been widely accepted for the formulation of answerable clinical questions. However, a study by [2] reported that not all clinical questions have all four PICO elements present. Two examples of questions maintained by the National Library of Medicine (NLM) [3] are "What is the best treatment for external otitis?" (Question 1) and "I have a lady with graves' disease (33 years old). She was trying to get pregnant when she was diagnosed with graves. So the question is, what is the best treatment for graves in someone who is trying to get pregnant, and if we use radioactive iodine, how long does she need to wait?" (Question 2). Both of the questions are

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 360–372, 2014.

categorized under "Treatment and Prevention". Question 1 represents a definitional question that contains only the P element ("external otitis"). Question 2 is described in paragraph format and contains both the P ("graves' disease", "lady") and I ("radioactive iodine") elements. An alternative intervention to the clinical condition and the expected treatment outcome, which denote the C and O elements respectively, are not stated in both Questions 1 and 2. As reported in [4], questions with the I/C and O elements are unlikely to go unanswered. Therefore, the visualization of PICO elements in documents relevant to a user's input query has the potential to assist the user in refining his/her information needs.

The paper presents a case study of utilizing similarity-based clustering to aid the visualization and exploration of interventions (i.e. the I and C elements) for the refinement of questions relating to treatments and drugs. The proposed user interface is illustrated in Fig. 1. As shown in the figure, the natural language (NL) question entered by the user contains only the P element ("breast cancer"). To assist the user in refining or clarifying his/her information needs, the user is allowed to explore a particular subject domain by browsing through the interventions which have been pre-clustered into a hierarchical structure. Simultaneously, the user can identify the interventions encompassed in each cluster and discover the relationships between the interventions. The most potent sets of PICO queries are produced by returning the interventions selected by the user (circled by black line in Fig. 1), accompanied by the P and O elements identified from the titles or abstracts. It is expected that through this process, a user can understand his/her information needs and obtain a more comprehensive knowledge about the domain of interest. An ambiguous query can also be refined by selecting the PICO query that best described the information needs.

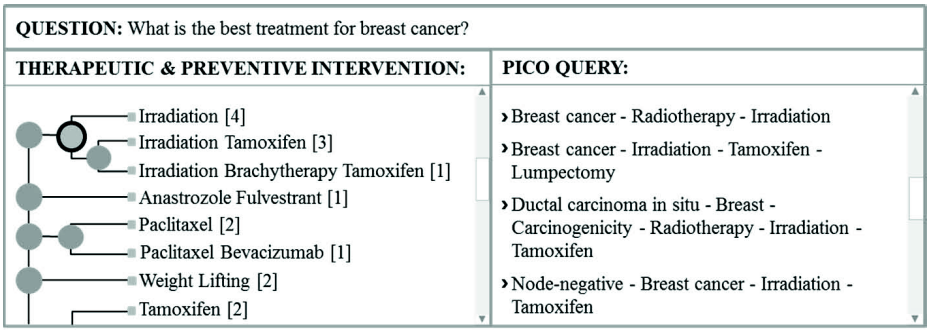

Fig. 1. The proposed user interface

2 Methodology

2.1 Collection of MEDLINE Documents

The processing of the NL question in Fig. 1 as described in Section 2.3 returns the medical concepts: "breast cancer" and "breast neoplasms". The concepts were used as the main search terms and the following filters were activated to retrieve relevant documents from the MEDLINE database: randomized controlled trial, abstract available,

publication date from 2002/10/01 to 2012/10/04, humans and English. The documents were limited to those published in 7 core journals: *N Engl J Med, JAMA, Ann Inter Med, Lancet, Br Med J, BMJ* and *BMJ (Clin Res Ed)*. The titles and abstracts of the documents were collected for the extraction of PICO elements.

2.2 Generation of PICO Sentences

Based on previous studies, the position of a sentence within an abstract is useful in determining the PICO elements that the sentence carries [5,6]. Two types of abstracts were identified: structured abstracts with internal section headings such as METHODS and RESULTS, and unstructured abstracts written in paragraph format without the headings. Both structured and unstructured abstracts were cut into three segments respectively based on the headings and the position of the sentences in the abstracts (Table 1). The extracted sentences were called in the remainder of this paper the "PICO sentences".

Table 1. Derivation of PICO sentences

Representation	Internal Section Heading	Position of Sentence
P	Introduction, Background, Objective	First 3 sentences
I/C	Method	Sentences in between the first and the last 3 sentences
O	Result, Conclusion	Last 3 sentences

2.3 Generation of PICO Elements

NL questions, titles and PICO sentences were processed by the MetaMap Transfer (MMTx) program [7] to semantically identify medical concepts as PICO elements. The program tokenizes the questions, titles and sentences into phrases, and returns a list of best matching concept candidates together with their associated semantic types from the Unified Medical Language System (UMLS) Metathesaurus. Each of the candidates was labeled with a concept unique identifier (CUI) number.

The concept candidates were post-processed using Rapidminer 5.2 [8] to identify the best matching candidates. Candidates with semantic types listed in Table 2 were recognized as PICO elements whereas those with other semantic types were deleted. Duplicate terms, synonyms and stopwords were removed by identifying their CUI numbers. For instance, "blood sugar" and "blood glucose" are synonyms with the same CUI number (i.e. C0005802). Examples of stopwords are "find", "release", "peer support", "still", "little" and "inform". If candidate terms of different lengths were identified at the same location in a document, candidates with the highest number of words were selected. For example, the processing of the phrase "management of orbital cellulitis" returns the concept candidates "orbital", "cellulitis" and "orbital cellulitis". "Orbital cellulitis" is selected and the rest are removed. For each document, a list of best matching medical concepts was collected respectively from the titles and the abstracts as PICO elements.

Table 2. Identification of PICO elements by semantic types (adapted from [6])

Representation	Semantic Type
P/O	Age group, Family group, Group, Human, Patient or disabled group, Population group, Acquired abnormality, Anatomical abnormality, Cell or molecular dysfunction, Congenital abnormality, Disease or syndrome, Experimental model of disease, Finding, Injury or poisoning, Mental or behavioral dysfunction, Neoplastic process, Pathologic function, Sign or symptom.
I/C	Daily or recreational activity, Amino acid, peptide, or protein, Antibiotic, Clinical drug, Eicosanoid, Enzyme, Hormone, Inorganic chemical, Lipid, Neuroreactive substance or biogenic amine, Nucleic acid, nucleoside, or nucleotide, Organic chemical, Organophosphorus compound, Pharmacologic substance, Receptor, Steroid, Vitamin, Diagnostic procedure, Therapeutic or preventive procedure.

2.4 Preprocessing of PICO Elements

The preprocessing involves three steps: (1) the I and C elements, i.e. the "interventions", were collected from titles, abstracts or a combination from both sections ("Tile + Abstract"), (2) the interventions were tokenized using "Loose" (LO) or "Strict" (ST) method, and (3) the interventions were weighted using normalized term frequency (TF), binary term occurrence (BI), term occurrence (TO) or term frequency-inverse document frequency (TF-IDF). The tokenization methods and weighting schemes are detailed as follow:

- LO: The interventions were tokenized into single term. For instance, the phrase "ascorbic acid" is tokenized into "ascorbic" and "acid"; ST: The interventions were not tokenized. For example, the phrase "breast radiotherapy" remains unchanged.
- TF: The ratio of the frequency of a term to the maximum term frequency of any term in a document, producing a numerical value between 0 and 1; BI: The occurrence of a term in a document with a binary value of 0 or 1; TO: A nominal value obtained by calculating the number of times a term occurs in a document; TF-IDF: A numerical value calculated by multiplying the frequency of a term in a document to the inverse of the number of documents in a collection that contains the term.

The three steps described above were achieved using Rapidminer 5.2 [8]. 24 sets of baseline data were generated based on the derivation of, and the tokenization methods and weighting schemes applied to the interventions.

2.5 Inter-document Similarity Tests

The baseline data were assembled into pairs of interventions. The similarity between each pair of interventions was computed using the "dist" and "simil" functions available in the R package "proxy" [9]. A total of 42 similarity/distance measures (Table 3) were utilized to compute the similarity or distance between the pairs of interventions. A distance measure was converted to a similarity measure using (1). The similarity values were normalized to a scale of 0 to 1. The normalized similarity value of each pair of interventions S_i was calculated using (2). S_{min} is the minimum similarity value and S_{max} is the maximum similarity value among all pairs of interventions.

Table 3. Similarity and distance measures

Data Type	Similarity Measure	Distance Measure
Numerical	Correlation, Cosine, eJaccard, fJaccard	Bhattacharyya, Bray, Canberra, Chord, Divergence, Euclidean, Geodesic, Hellinger, Manhattan, Soergel, Supremum, Wave, Whittaker
Binary	Braun-blanquet, Dice, Fager, Faith, Hamman, Jaccard, Kulczynski1, Kulczynski2, Michael, Mountford, Mozley, Ochiai, Phi, Russel, Simple Matching, Simpson, Stiles, Tanimoto, Yule, Yule2	-
Nominal	Chi-squared, Cramer, Pearson, Phi-square, Tschuprow	-

$$Similarity = \frac{1}{Distance + 1} \quad (1) \qquad S_{norm} = \frac{S_i - S_{min}}{S_{max} - S_{min}} \quad (2)$$

A retrospective analysis was conducted manually to judge the similarity of the pairs of interventions. Interventions which are highly similar were identified as paired interventions whereas those with low similarity were identified as unpaired interventions. Histograms and boxplots were created to assess the effectiveness of the similarity/distance measures in separating paired interventions from unpaired interventions. A one-way ANOVA was performed to compare the means of similarity values between paired and unpaired interventions. The mean difference (MD) was calculated to compare the performance of the 42 measures on the 24 sets of baseline data using (3). $\overline{S_{paired}}$ is the mean of similarity values of paired interventions and $\overline{S_{unpaired}}$ is the mean of similarity values of unpaired interventions.

$$Mean\ Difference\ (MD) = \overline{S_{paired}} - \overline{S_{unpaired}} \quad (3)$$

2.6 Cluster Structure Analysis

Similar interventions were clustered together using agglomerative hierarchical clustering methods: average-link (AL), complete-link (CL) and ward-link (WL). The three types of clusterings were generated using the "hclust" function available in the R package "stats" [9].

A sample of clustering was shown in Fig. 2a. Based on the figure, each interventions (e.g. "Tamoxifen Bevacizumab") represents a single document. The number of levels and the number of documents that a user will need to explore to obtain all the relevant documents for a topic of interest were identified. For instance, a user will need to explore two levels to discover two documents representing the topic "Irradiation Tamoxifen".

The precision (P), recall (R) and F-measure (F) of each cluster were calculated. P is the ratio of relevant documents retrieved for a given topic (N_{Rel}) over the total number of relevant and irrelevant documents retrieved ($N_{Rel} + N_{Irrel}$) (4). R is the ratio of relevant documents retrieved for a given topic (N_{Rel}) over the total number of relevant documents retrieved and not retrieved ($N_{Rel} + M_{Rel}$) (5). F is the harmonic mean of P and R (6).

$$Precision\ (P) = \frac{N_{Rel}}{N_{Rel} + N_{Irrel}} \quad (4) \qquad Recall\ (R) = \frac{N_{Rel}}{N_{Rel} + M_{Rel}} \quad (5)$$

$$F - measure\ (F) = 2 \times \frac{P \times R}{P + R} \quad (6)$$

Random-baseline clusterings were constructed to evaluate the information retrieval performance of similarity-based clusterings. A random-baseline clustering was created by randomly assigning the interventions into a clustering that has the same number of clusters and the same number of documents in each cluster of a similarity-based clustering. An example is given in Fig. 2. The P, R and mean average precision (MAP) over 25 topics were computed using the TREC_EVAL program [10].

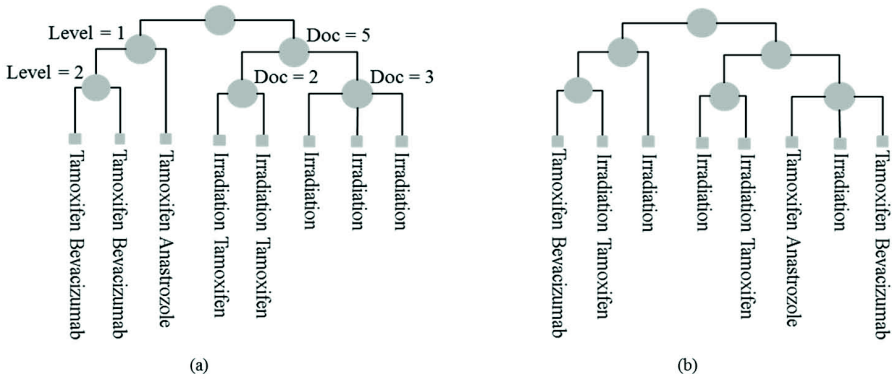

Fig. 2. (a) Similarity-based clustering and (b) random-baseline clustering. Doc = number of documents

3 Results

3.1 The Inter-document Similarity Tests

42 types of similarity/distance measures were employed to calculate the similarities between pairs of interventions. A value close to 1 indicates strong similarity whereas a value close to 0 means low similarity. The MD was calculated to indicate the difference between the mean of paired and the mean of unpaired similarities. The larger the MD, the greater the differentiation and the less overlap between the distributions of paired and unpaired similarities. Table 4 summarizes the measures that produced the highest MD over the 24 sets of baseline data. The table revealed that: (1) BI is better than other weighting schemes, (2) the tokenization method, LO, is superior to ST, (3) interventions derived from title are better than those derived from abstract or "title + abstract", and (4) Yule gives the highest MD compared to other measures.

One of the most popular distance measures between two document vectors is the Cosine similarity. Fig. 3a compares the MD of Yule to the MD of Cosine with an

increase in number of pairs of interventions. The figure shows that the number of pairs has little influence on the performance of Yule and Cosine. The MD of Yule (average MD = 0.86 ± 0.04) is evidently higher than the MD of Cosine (average MD = 0.50 ± 0.02 and 0.40 ± 0.02 respectively when tokenized by TF and TF-IDF). For both measures, a significant difference in similarity values between paired and unpaired interventions was found ($p < 0.005$).

Table 4. Similarity/distance measures that produced the highest mean difference (MD)

Weighting Scheme	Tokenization Method	Derivation of Interventions		
		Title (MD)	Abstract (MD)	Title + Abstract (MD)
TF	LO	eJaccard (0.53)	Cosine (0.34)	Cosine (0.45)
	ST	eJaccard (0.44)	Cosine (0.28)	Cosine (0.36)
TF-IDF	LO	eJaccard (0.43)	Cosine (0.24)	Cosine (0.31)
	ST	eJaccard (0.36)	Cosine (0.20)	Cosine (0.27)
BI	LO	Yule (0.86)	Yule (0.67)	Yule (0.74)
	ST	Yule (0.66)	Yule (0.53)	Yule (0.63)
TO	LO	Pearson (0.57)	Pearson (0.34)	Pearson (0.53)
	ST	Pearson (0.49)	Pearson (0.31)	Pearson (0.43)

Histograms and boxplots (Fig. 3b) were plotted to investigate the frequency distribution of similarity values of 450 paired and 450 unpaired interventions. As shown in the figure, the less overlap between the two histograms, the greater the separation between the two distributions. The paired histogram for BI-Yule combination skewed significantly to the right, showing that most of the paired interventions has a similarity value close or equal to 1. In contrast, the paired histograms for the two Cosine combinations are relatively flat with similarity values range between 0 and 1 (as indicated also by the whiskers and outliers in boxplots). The degree of overlap for the three combinations looks apparently the same. In terms of classifiability, Yule gave a more clear-cut separation of paired and unpaired similarities in histograms and boxplots than Cosine.

In summary, the similarity measure, Yule, performed better than other measures at identifying paired interventions or at differentiating between paired and unpaired interventions derived from titles and which had been weighted and tokenized respectively using BI and the LO method.

3.2 The Clustering Tests

The similarities between the interventions that occurred in a collection of 114 documents were calculated using the BI-LO-Title-Yule combination. Hierarchical clusterings were computed using AL, CL and WL algorithms. As shown in Fig. 4a (1st column), the structure of AL and CL clusterings are wide with many branches at the top of the hierarchies, whereas for WL clustering, branches are located mainly at the bottom of the hierarchy. The CL algorithm produced a structure with lower number of levels (the highest number of levels = 5 compared to 10 for WL and 15 for AL). A comparison of the structures obtained by calculating the similarities using the TF-LO-Title-Cosine combination (Fig. 4a, 2nd column) revealed a higher number of levels in the clusterings (the highest number of levels = 7 for CL, compared to 13 for WL and

18 for AL). The higher the number of levels, the longer it takes for a user to browse and search for a topic in a hierarchy. The Yule-based CL clustering, compared to other clusterings, provides a better structure in terms of the number of levels that a user needs to explore to reach a topic of interest.

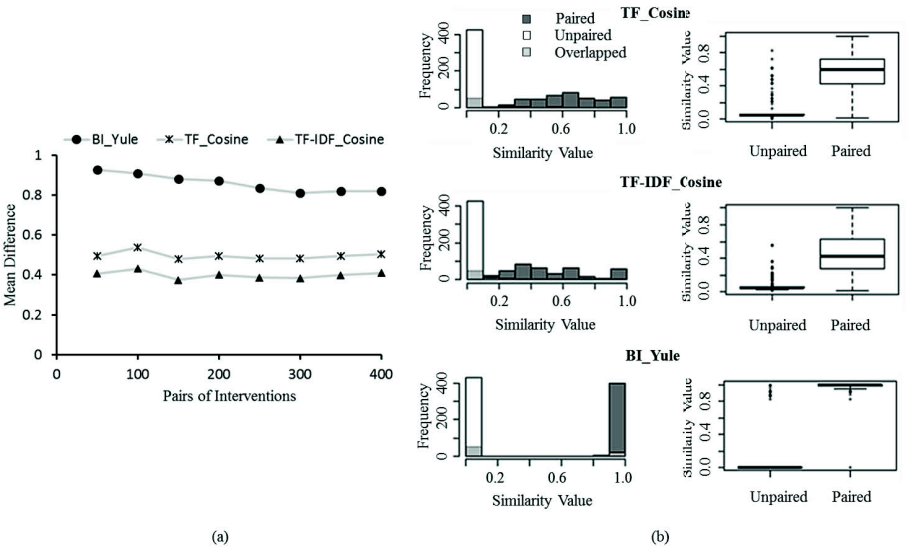

Fig. 3. (a) Mean difference against number of pairs of interventions; (b) Histograms and box-plots of the distribution of paired and unpaired similarities. Derivation of interventions: title, tokenization method: LO

The clusters that best represent 33 topics that covered in the 114 documents were identified by computing the P, R and F of each cluster in a hierarchy. Table 5 shows the level of the clustering hierarchy (Lev), the number of documents (Doc), the number of relevant documents (Rel) and the P, R and F values of a cluster in a Yule-based CL clustering. A good cluster is supposed to contain as many relevant documents as possible with high P and high R. The F-measure quantifies the balance between P and R. The higher the F value, the higher the quality of a cluster. It can be seen from Table 5 that: (1) relevant documents are grouped in one (e.g. Level 1 of Topic 1, R = 1.0) or two clusters (e.g. Level 1 of Topic 2, R = 0.5 and 0.5 respectively), (2) the best clusters appear at the top of the structure with high P, R and F for Topics 1, 2 and 4, (3) the best clusters occur at the bottom of the structure with high P, R and F for Topic 3, (4) Topic 4 can be identified without exploring the structure (Level = 0), and (5) some of the relevant documents are grouped in different clusters with irrelevant documents (e.g. Level 1 of Topic 3, P = 0.6 and 0.3 respectively). The results indicate that the best clusters located at different levels of the structure.

Table 6 shows the average number of levels that a user needs to explore to discover the best clusters for the 33 topics. The Yule-based CL clustering provides the best hierarchical structure for the exploration of different topics, followed by the

Cosine-based CL clustering (Average No. of Level = 1.70 ± 1.10 and 2.33 ± 1.95 respectively). The findings suggest that the best clusters appear on average at the top two levels of a CL clustering. This was evaluated by identifying the clusters with the highest F-measure (F_{Max}) from the top two levels. The percentages of relevant documents covered by the clusters were then calculated and are shown in Table 7. On average over the 33 topics, the clusters from the top two levels contain approximately 81% and 79% of the relevant documents respectively for Yule-based and Cosine-based CL clusterings. This suggests that only a small number of clusters that will need to be further explored to obtain all the relevant documents from the clusterings.

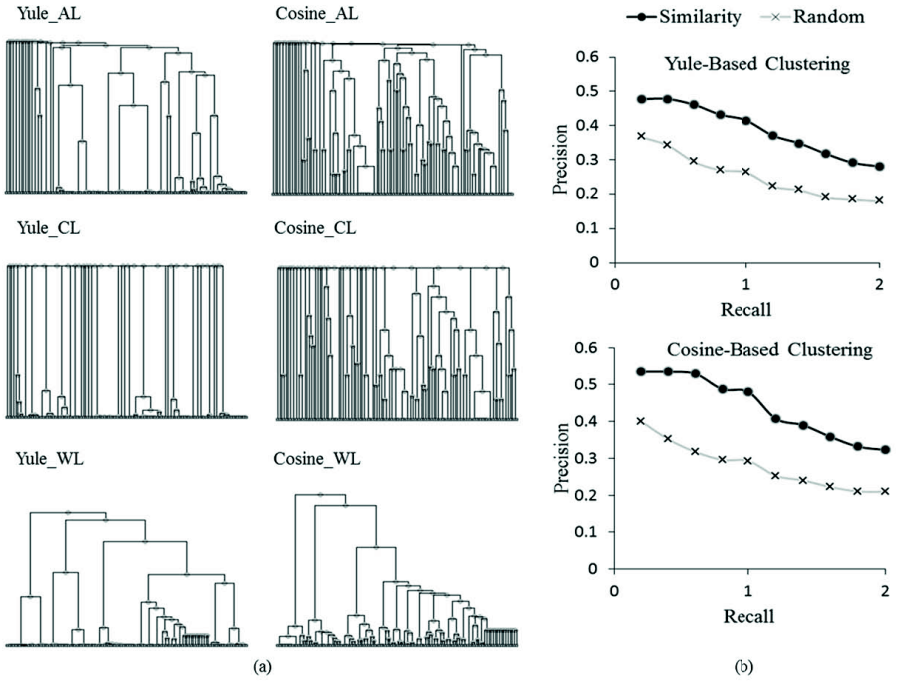

Fig. 4. (a) A comparison of clusterings by similarity measures and clustering methods; (b) The precision-recall performance of similarity-based and random-baseline CL clusterings

The effectiveness of similarity-based clusterings in grouping similar interventions to the same or small number of clusters were evaluated by comparing with random-baseline clusterings. Interventions were grouped into different clusters without similarity constraint to produce a random-baseline clustering. A total of 25 topics were created for the evaluation. Each topic was treated as a query. Similarity-based clusterings outperform random-baseline clusterings in terms of mean average precision (MAP = 0.43 vs. 0.24 and 0.48 vs. 0.25 respectively for Yule-based and Cosine-based CL clusterings). This is further indicated in the P-R curves shown in Fig. 4b.

The overall clustering results indicate that the top two levels of a Yule-based CL clustering provide the most appropriate hierarchical clustering for the exploration and visualization of interventions.

Table 5. Examples of the distribution of the best cluster in a Yule-based CL clustering

Topic	Lev	Doc	Rel	P	R	F
1	**1**	**10**	**10**	**1.0**	**1.0**	**1.0**
	2	2	2	1.0	0.2	0.3
	2	2	8	1.0	0.8	0.9
	3	3	7	1.0	0.7	0.8
2	1	15	4	0.3	0.5	0.4
	1	**4**	**4**	**1.0**	**0.5**	**0.7**
	⋮	⋮	⋮	⋮	⋮	⋮
	4	3	1	0.3	0.1	0.2
3	1	6	1	0.6	0.3	0.2
	1	11	3	0.3	0.8	0.4
	2	2	1	0.5	0.3	0.3
	2	10	3	0.3	0.8	0.4
	3	**3**	**3**	**1.0**	**0.8**	**0.9**
4	**0**	**1**	**1**	**1.0**	**0.5**	**0.5**
	0	**1**	**1**	**1.0**	**0.5**	**0.5**

Table 6. Location of the best cluster by the average number of level over 33 topics

Clustering Method	Similarity Measure	Average ± SD No. of Level
CL	Cosine	2.33 ± 1.95
CL	Yule	1.70 ± 1.10
AL	Cosine	10.12 ± 4.84
AL	Yule	8.21 ± 3.56
WL	Cosine	7.67 ± 3.91
WL	Yule	5.97 ± 2.49

Table 7. Percentages of relevant documents in top clusters in CL clusterings

	Cosine_CL				Yule_CL			
Topic	F_{Max}	Doc	Rel	%	F_{Max}	Doc	Rel	%
Q1	0.6	8	11	73	0.6	8	11	73
Q2	0.1	1	2	50	0.2	1	2	50
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Q33	0.6	2	2	100	0.6	2	2	100
Mean ± SD				79 ± 23				81 ± 23

4 Discussion and Conclusions

What types of clinical information do doctors need? Where do they search for information? An early study by [11] reported that approximately 33% of information needs related to treatment of specific conditions, 25% to diagnosis and 14% to drugs. Similar findings were reported in [12] that the top categories of information needs were treatment/therapy (38%), diagnosis (24%) and drug therapy/ information (11%). Studies by [13,14] further supported that one of the doctors' greatest information needs is for information about treatments and drugs. The primary electronic resource used by doctors for evidence-based clinical decision making is MEDLINE [15,16]. Junior doctors accessed MEDLINE (44%), UpToDate (42%), internet search engines (5%), MDCONSULT (3%) and the Cochrane Library (2%) for clinical information [17]. The findings support the use of MEDLINE in this study as the preferred information source for PICO elements.

The Use of PICO Framework for Query Refinement. One of the obstacles that prevents physicians from answering patient-care question is the tendency to formulate

unanswerable question [1]. To formulate an answerable question, physicians are recommended to change their search strategies by rephrasing their questions [17]. Other studies recommended the use of question frameworks such as PICO, PICOT, PICOS and PESICO for the formulation of clinical question [18,19,20,21]. A study evaluating the use of PICO as a knowledge representation reported that the framework is primarily centered on therapy question [2]. This supports the focus of the present study on refining questions relating to treatments and drugs. An earlier study by [4] found that questions that contain a proposed intervention and a relevant outcome were unlikely to go unanswered. It is recommended by [22] that at least 3 of the PICO elements are needed to formulate an answerable question. In summary, the completeness of PICO elements determines whether a clinical question is likely to be answered.

Visualization of Interventions Using Similarity-Based Clustering. Current medical QA (MedQA) systems focus on providing direct and precise answers to doctors' questions. A recent review by [23] concluded that current MedQA systems have limitations in terms of the types and formats of questions that they can process. The InfoBot [24] and the EpoCare [25] systems can only handle structured queries in PICO format but not in NL. An example of PICO query is "*Atrial Fibrillation AND Warfarin AND Aspirin AND Secondary Stroke*". The use of the system may be limited by the ability of users to apply Boolean operators (e.g. AND and OR) and by the lack of vocabulary due to limited knowledge of a particular domain. The AskHermes system [26,27], on the other hand, accepts both well-structured and ill-formed NL questions. For example, "*What is the best treatment for a needle stick injury after a human immunodeficiency virus exposure?*" A poorly formulated question cannot be refined. This can in turn lead to the discovery of irrelevant documents. Current MedQA systems assume that users are aware of their knowledge deficit. Little research has focused on assisting users in formulating high quality questions, supporting them in exploring a problem domain and clarifying their information needs.

The present study adopted the concept of system-mediated information access, introduced by [28], to assist users in refining an ill-defined question. It is expected that users can clarify or refine their information needs through browsing and searching interventions which have been pre-clustered into a hierarchical structure. The inter-document similarity and cluster structure analysis revealed that the combination of BI-LO-Title-Yule-CL produced the most appropriate hierarchical clustering for the visualization of interventions. The Yule measure appeared to be slightly better than the Cosine measure at contributing to the identification of similar interventions. The Cosine similarity, which measures the cosine of the angle between two vectors, has been applied to both document clustering [29] and short text clustering [30]. The Yule similarity calculates the strength of association between binary variables. Though not as well studied as the Cosine similarity, [31] reported an improvement in clustering performance using the Yule measure. The cluster structure analysis revealed that documents with similar interventions are likely to be grouped into the same cluster. The top two levels of a CL clustering provide the most appropriate structure for the exploration of different topics. Previous work by [32] reported that AL produces a more effective clustering than CL for information retrieval. However, in the present study, the AL clustering requires users to explore a higher number of levels to discover a problem domain. Doctors often have very tight schedules. When seeking information for patient care, they are more likely to look for

information that can be accessed quickly with minimal effort [11]. Therefore, it is argued that CL provides a quicker and more appropriate clustering than AL for the visualization and exploration of interventions.

Limitations. The study was conducted only on MEDLINE articles relevant to a single question. The single source of documents may restrict the applicability of the findings from this study to documents from other resources such as the Cochrane Library and UpToDate. The question tested was posed with only the P element. Further analysis should be undertaken with higher number of questions addressed with different combinations of PICO elements. Compared to the titles, a higher number of interventions were collected from the abstracts. The case study however shows that the title-based approach superior to the abstract-based approach. The study can be improved by evaluating the effectiveness of the methodologies used for PICO extraction and the effects of different numbers of interventions between two documents on the measurement of similarity. Despite of the limitations, the experimental results show that the similarity-based clustering approach has the potential to aid the visualization and exploration of interventions for the applications of clinical information needs clarification and query refinement.

References

1. Ely, J.W., Osheroff, J.A., Chambliss, M.L., Ebell, M.H., Rosenbaum, M.E.: Answering physicians' clinical questions: obstacles and potential solutions. J. Am. Med. Inform. Assoc. 12(2), 217–224 (2005)
2. Huang, X., Lin, J., Demner-Fushman, D.: Evaluation of PICO as a knowledge representation for clinical questions. In: AMIA Annual Symposium Proceedings 2006, pp. 359–363 (2006)
3. Cao, Y., Liu, F., Simpson, P., Antineau, L., Bennet, A., Cimino, J.J., Ely, J., Yu, H.: AskHERMES: an online question answering system for complex clinical questions. J. Biomed. Inform. 44(2), 227–288 (2011)
4. Bergus, G.R., Randall, C.S., Sinift, S.D., Rosenthal, D.M.: Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? Arch. Fam. Med. 9(6), 541–547 (2000)
5. Demner-Fushman, D., Lin, J.: Answering clinical questions with knowledge-based and statistical techniques. Association for Computational Linguistics 33(1), 63–103 (2007)
6. Boudin, F., Nie, J.Y., Bartlett, J.C., Grad, R., Pluye, P., Dawes, M.: Combining classifiers for robust PICO element detection. BMC Med. Inform. Decis. Mak. 10(1), 29–36 (2010)
7. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of AMIA Annual Symposium 2001, pp. 17–21 (2001)
8. RapidMiner: Report the future, `http://rapid-i.com/content/blogcategory/38/69/` (Assessed: August 2013)
9. R: The R project for statistical computing, `http://www.r-project.org` (Assessed: August 2013)
10. Trec_eval, `http://trec.nist.gov/trec_eval/` (Assessed: August 2013)
11. Smith, R.: What clinical information do doctors need? BMJ 313(7064), 1062–1068 (1996)
12. Davies, K., Harrison, J.: The information-seeking behavior of doctors: a review of the evidence. Health Info. Libr. J. 24(2), 78–94 (2007)
13. Schwartz, K., Northrup, J., Crowell, K., Lauder, N., Neale, A.V.: Use of on-line evidence-based resources at the point of care. Family Medicine 35(4), 251–256 (2003)

14. Yu, H., Cao, Y.G.: Automatically extracting information needs from ad hoc clinical questions. In: AMIA Annual Symposium Proceedings 2008, pp. 96–100 (2008)
15. Davies, K.: UK doctors awareness and use of specified electronic evidence-based medicine resources. Inform. Health Soc. Care 36(1), 1–19 (2011)
16. Schilling, L.M., Steiner, J.F., Lundahl, K., Anderson, R.J.: Residents' patient-specific clinical questions: opportunities for evidence-based learning. Academic Medicine 80(1), 51–56 (2005)
17. Ely, J.W., Osheroff, J.A., Maviglia, S.M., Rosenbaum, M.E.: Patient-care questions that physicians are unable to answer. J. Am. Med. Inform. Assoc. 14(4), 407–414 (2007)
18. Schardt, C., Adams, M.B., Owens, T., Keitz, S., Fontelo, P.: Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med. Inform. Decis. Mak. 7(1), 16 (2007)
19. Rios, L.P., Ye, C., Thabane, L.: Association between framing of the research question using the PICOT format and reporting quality of randomized controlled trials. BMC Med. Res. Methodol. 10(1), 11–18 (2010)
20. Robinson, K.A., Saldanha, I.J., Mckoy, N.A.: Frameworks for determining research gaps during systematic reviews. In: Methods Future Research Needs Reports, vol. 2. Agency for Healthcare Research and Quality, Rockville (MD) (2011)
21. Schlosser, R.W., Koul, R., Costello, J.: Asking well-built questions for evidence-based practice in augmentative and alternative communication. Journal of Communication Disorders 40(3), 225–238 (2007)
22. Staunton, M.: Evidence-based radiology: steps 1 and 2 – asking answerable questions and searching for evidence. Radiology 242(1), 23–31 (2007)
23. Athenikos, S.J., Han, H.: Biomedical question answering: a survey. Comput. Methods Programs Biomed. 99(1), 1–24 (2010)
24. Demner-Fushman, D., Seckman, C., Fisher, C., Hauser, S.E., Clayton, J., Thoma, G.R.: A prototype system to support evidence-based practice. In: AMIA Annual Symposium Proceedings 2008, pp. 151–155 (2008)
25. Niu, Y., Hirst, G., McArthur, G., Rodriguez-Gianolli, P.: Answering clinical questions with role identification. In: Proceedings, Workshop on Natural Language Processing in Biomedicine, 41st Annual Meeting of the Association for Computational Linguistics, pp. 73–80 (2003)
26. Yu, H., Kaufman, D.: A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. In: Pacific Symposium on Biocomputing 2007, pp. 328–339 (2007)
27. Cao, Y.G., Cimino, J.J., Ely, J., Yu, H.: Automatically extracting information needs from complex clinical questions. J. Biomed. Inform. 43(6), 962–971 (2010)
28. Muresan, G., Harper, D.J.: Topic modelling for mediated access to very large document collections. J. Am. Soc. Inf. Sci. Technol. 55(10), 892–910 (2004)
29. Subhashini, R., Kumar, V.: Evaluating the performance of similarity measures used in document clustering and information retrieval. In: Proceedings of the First International Conference on Integrated Intelligent Computing 2010, pp. 27–31 (2010)
30. Rangrej, A., Kulkarni, S., Tendulkar, A.: Comparative study of clustering techniques for short text documents. In: Proceedings of the 20th International Conference Comparison on World Wide Web 2011, pp. 111–112 (2011)
31. Malik, H.H., Kender, J.R.: High quality, efficient hierarchical document clustering using closed interesting itemsets. In: Proceedings of the Sixth International Conference on Data Mining 2006, pp. 991–996 (2006)
32. Aljaber, B., Stokes, N., Bailey, J., Pei, J.: Document clustering of scientific texts using citation contexts. Information Retrieval 13(2), 101–131 (2010)