Journal of Bioinformatics and Computational Biology

MeSHSim: An R/Bioconductor package for measuring semantic similarity over MeSH headings and MEDLINE documents

JING ZHOU [1,2], YUXUAN SHUI [1,2], SHENGWEN PENG [1,2], XUHUI LI [3], HIROSHI MAMITSUKA [4], SHANFENG ZHU [1,2*]

[1] *School of Computer Science, Fudan University*
Shanghai, China 200433
[2] *Shanghai Key Laboratory of Intelligent Information Processing, Fudan University*
Shanghai, China 200433
[3] *School of Information Management, Wuhan University*
Wuhan, China 430072
[4] *Bioinformatics Center, Institute for Chemical Research, Kyoto University*
Kyoto, Japan 611-0011
[1] *jingzhou12@fudan.edu.cn*
[1] *10300240045@fudan.edu.cn*
[1] *swpeng14@fudan.edu.cn*
[1] *zhusf@fudan.edu.cn*
[2] *jingzhou12@fudan.edu.cn*
[2] *10300240045@fudan.edu.cn*
[2] *swpeng14@fudan.edu.cn*
[2] *zhusf@fudan.edu.cn*
[3] *lixuhui@whu.edu.cn*
[4] *mami@kuicr.kyoto-u.ac.jp*

Received (September 2, 2015)
Revised (Day Month Year)
Accepted (Day Month Year)

Currently all MEDLINE documents are indexed by Medical Subject Headings (MeSH). Computing semantic similarity between two MeSH headings as well as two documents has become very important for many biomedical text mining applications. We develop an R package, MeSHSim, which can compute nine similarity measures between MeSH nodes, by which similarity between MeSH Headings as well as MEDLINE documents can be easily computed. Also MeSHSim supports querying hierarchy information of a MeSH heading and retrieving MeSH headings of a query document, and can be easily integrated into pipelines for any biomedical text analysis tasks. MeSHSim is released under GPL (General Public License), and available through Bioconductor and from Github at `https://github.com/JingZhou2015/MeSHSim`

Keywords: MeSH, Semantic Similarity, MEDLINE documents, R/Bioconductor Package

1. Introduction

MeSH (Medical Subject Headings) is a controlled vocabulary used by National Library of Medicine to index MEDLINE documents. MeSHSim consists of a set of de-

*To whom correspondence should be addressed

scription terms, which are organized in a hierarchical structure (called MeSH trees), where more general terms appear at nodes closer to the root and more specific terms appear at nodes closer to leaves [18]. Each MEDLINE document is manually annotated with a set of (usually 10-15) MeSH headings, including around three to five major headings, representing main topics of the corresponding document. Computing semantic similarities between two MeSH headings as well as two documents (one document having a set of MeSH headings) has been proved very useful to improve the performance of many biomedical text mining tasks. The semantic similarity between two MeSH headings is utilized to improve information retrieval by ranking documents retrieved from MEDLINE [20]. The semantic similarity is incorporated into the citation searching process through re-ranking the top N documents based on the similarity calculated using MeSH headings [3]. A novel semantic similarity based information retrieval model was proposed by Hliaoutakis *et al.* [9] that is able to find relevant documents with similar headings by query expansion. Furthermore, semantic similarity between MeSH headings was used to initialize clustering and achieved promising clustering results [26]. It is also applied to MEDLINE document clustering and achieved promising results [28,7].

The existing techniques to measure similarity between two MeSH headings can generally be divided into two categories: path-based measures and information content (IC)-based measures. As the specificity of a MeSH heading is usually determined by its location in the MeSH tree, path-based measures compute semantic similarity between two MeSH headings as a function of the path linking the headings. On the other hand, the IC-based measures rely on the frequencies of two MeSH headings involved to measure their difference.

As such, MeSH semantic similarity is widely used and have been well studied, but there have been no available tools for computing similarity between MeSH headings and also MEDLINE documents, except an online tool THE MESH SIMILARITY (http://sce.uhcl.edu/biomedsim/). This tool (last updated in 2011) cannot be used for computing similarity between two documents. Importantly, this is a web server, which cannot be a building block of a text mining software on MEDLINE documents. In this light, we provide an R package, MeSHSim, to compute semantic similarity among MeSH headings and also MEDLINE documents. MeSHSim can be easily integrated into biomedical text mining applications, which will be built by users.

The rest of the paper is organized as follows: Section 2 describes nine semantic similarity measures implemented in the MeSHSim package, including five path-based measures and four IC-based measures; Section 3 shows utilities and examples of nine functions integrated in MeSHSim. Section 4 presents two case studies using MeSHSim in a real scenario. Finally, we give a brief discussion on the MeSHSim package and conclude the paper in Section 5.

2. Implementation

We first explain many similarity measures, which have been proposed so far. Generally we can divide these measures into two types: path-based and information content (IC)-based measures [2]. In our package we implement five path-based and four IC-based similarity measures. In this section we will further explain the way to compute semantic similarities between two MeSH Headings and then that between two documents (MeSH Heading Sets). Finally we will briefly mention required packages for MeSHSim

2.1. *Similarity Measures*

2.1.1. *Path-based Similarity Measure*

This kind of measurement is based on spread activation theory proposed by Cohen *et al.* [5], which assumes that the conceptual network is organized according to semantic similarity, in which concepts are assumed to be represented as nodes. As all the headings of the ontology are organized in a hierarchy, where more general headings are near the root of the hierarchy, and more specific ones near at the leaves, it is convenient to measure similarity as a function of the length of the path linking the headings and on the position of the headings in the hierarchy. Most of the measures using the hierarchy structure of ontology are actually based on: 1) path length (i.e., shortest path length/distance between the two heading nodes) and 2) depth of heading nodes in the ontology.

i. SP: Shortest Path [4]. The measurement is motivated by the observation that the distance between two nodes is often proportional to the number of edges separating the two nodes in the hierarchy. This measure is designed to find the gap between the local path length and the maximum path length, and use as the semantic score.

$$Sim_{SP} = (MAX - L)/MAX \tag{1}$$

where MAX is the maximum path length between two headings in the hierarchy, L is the shortest path between two headings.

ii. WL: Weighted Links [22]. This measure extends the Shortest Path measure by introducing the weighted edges in counting the path length.

$$Sim_{WL} = \frac{WMAX - WL}{WMAX} \tag{2}$$

where $WMAX = max_{i,j} WL_{ij}$ is the maximum weighted path length, and

$$WL_{ij} = \sum_{k \in path_{ij}} \frac{1}{H_k} \tag{3}$$

where H_k is the depth of node k in the hierarchy

iii. WP: headingual similarity [25]. This measure is designed to find the nearest common ancestor of the two headings. The path length from this ancestor heading

to the root of the ontology is scaled by the sum of path length of the two headings.

$$Sim_{WP} = \frac{2H_c}{H_1 + H_2} \tag{4}$$

where c is the nearest common ancestor of the two headings.

iv. LC: Leacock and Chodorow [13]. This measure is to scale the shortest path by twice the maximum depth of the hierarchy.

$$Sim_{LC} = 1 - \frac{log(1+L)}{1+2D} \tag{5}$$

where D is the maximum depth of the heading in the ontology.

v. Li: Li et al [14]. The measure combines the shortest path and the depth of the closest common ancestor in a non-linear function.

$$Sim_{Li} = e^{-\alpha L}\frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}} \tag{6}$$

where α and β are parameters scaling the contribution of shortest path length and depth respectively. H is the minimum depth of the nearest common ancestor.

2.1.2. *Information-Content based Similarity Measure*

The notion of information content of the heading practically has to do with the frequency of the heading in a given document collection. Given a corpus C, $p(c)$ is the probability of encountering an instance of heading c. The heading probability is defined as $p(c) = freq(c)/N$, where N is the total number of headings that appear in C, $freq(c)$ corresponds to the frequency of heading c. Additionally, the frequency counts of every heading includes the frequency counts of subsumed headings in an IS-A hierarchy. Then, the information content of c can be computed: $I(c) = -\log p(c)$.

i. Lord: Lord [16]. The first way to compare two headings is by using a measure that simply uses the probability of nearest common ancestor.

$$Sim_{lord} = 1 - p(c) \tag{7}$$

where c is the nearest common ancestor of heading c_1 and c_2.

ii. Resnik: Resnik [21]. This measure signifies that the more information two headings share in common, the more similar they are.

$$Sim_{Resnik} = I(c) \tag{8}$$

iii. Lin: Lin [15]. This measure is the same as WP, except that the information content is used, instead of node depth.

$$Sim_{Lin} = \frac{2 * I(c)}{I(c_1) + I(c_2)} \tag{9}$$

iv. JC: Jiang and Conrath [11]. The measure defined a distance function as follows,

$$Dist_{JC} = I(c_1) + I(c) - 2 * I(c) \tag{10}$$

We use an exponential function to transform the distance into a similarity with constant λ. A large λ will yield a high similarity value even for weakly related headings.

$$Sim_{JC} = e^{-\frac{Dist_{JC}(c_1,c_2)}{\lambda}} \tag{11}$$

2.2. *Semantic Similarity between MeSH Headings*

As not all of the MeSH headings are represented by only one tree node, two frameworks have been proposed to compute the semantic similarity between two MeSH headings: node-based framework first proposed by Zhu *et al.* [28] and heading-based framework. On the one hand, node-based framework implies that we project the two MeSH main headings onto the tree structure and calculate the similarity between the two projected node sets. On the other hand, heading-based framework tries to build a relation structure among main headings through the MeSH tree structure and then only consider the nearest path between them. Please note that the two frameworks differ from each other in computing the information content.

2.2.1. *Node-based Framework*

Node-based framework uses the Average Maximum Match method proposed by [24]. Considering a general case in which each MeSH main heading has one or multiple tree nodes, for each MeSH nodes v in main heading M, the maximum similarity between v and any MeSH nodes in M' is used to represent its contribution to the similarity between M and M':

$$\mathrm{Sim}(M, M') = \frac{\sum_{v \in M} \max_{v' \in M'} \mathrm{Sim}(v, v') + \sum_{v' \in M'} \max_{v \in M} \mathrm{Sim}(v, v')}{|M| + |M'|},$$

where $|M|$ indicates the number of MeSH node in M.

Computing IC: As each MeSH main heading corresponds to one or several MeSH tree nodes, accordingly the frequency of these related tree nodes and their ancestor nodes should be updated simultaneously. Then, the total number N is defined as the frequency of global root node. Finally, the information content of each node is computed as previous mentioned.

2.2.2. *Heading-based Framework*

Heading-based framework treat each MeSH main heading as a basic computational element, however many headings could be mapped to not a single postion on the tree structure; so when projected to the tree structure, there might be several position-postion relationship for a MeSH heading pair and we can calculate several candidate similarity scores. Typically, the heading-based similarity is computed by as simpler idea given as follows [19]

$$\mathrm{Sim}(M, M') = \max_{v \in M, v' \in M'} \mathrm{Sim}(v, v').$$

Table 1. Nine functions in MeSHSim and their input & output

name	input	output
nodeSim	two MeSH nodes	similarity
headingSim	two MeSH headings	similarity
headingSetSim	two MeSH heading sets	similarity
docSim	two MEDLINE documents	similarity
mnodeSim	multiple MeSH nodes	similarities
mheadingSim	multiple MeSH headings	similarities
nodeInfo	MeSH node	tree information
termInfo	MeSH heading	tree information
docInfo	MEDLINE document	document information

Computing IC: As each MeSH main heading could be mapped to the tree structure, the frequency of it and its ancestor main headings (through its projection onto the tree structure) should accumulate simultaneously. And the total number N is denoted as the sum of the frequencies of all the main headings. Therefore, the information content of each main heading is computed as previous mentioned.

2.3. *Similarity between Two Documents (MeSH Heading Sets)*

As each MEDLINE article is marked by a set of MeSH headings, the similarity between two documents can be measured by the similarity between two MeSH heading sets, which relate to the two documents. A measurement is proposed by Zhu *et al.* [28] to calculate sematic similarity between two MeSH sets, which adopt the idea of Average Maximum Match(AMM). Given two documents, D and D', the similarity between two MeSH headings can be calculated as follows:

$$\mathrm{Sim}(D, D') = \frac{\sum_{M \in D} \max_{M' \in D'} \mathrm{Sim}(M, M') + \sum_{M' \in D'} \max_{M \in D} \mathrm{Sim}(M, M')}{|D| + |D'|},$$

where $|D|$ indicates the number of MeSH headings in document D.

2.4. *Required packages*

MeSHSim needs three R packages: bitops, XML and RCurl, where bitops, used by RCurl, supports bitwise operations of integer vectors, XML supports reading XML documents and RCurl [12] is to fetch document information from PubMed. They are freely available at CRAN (Comprehensive R Archive Network).

3. Functions and examples

Table 1 shows nine functions implemented in MeSHSim. The first four functions compute pairwise similarities, which take a value between zero and one with higher

value being more similar. For example, *nodeSim* (the default parameter of "method" is SP, standing for Shortest Path), *headingSim* and *headingSetSim* (the default parameter of "frame" is "node", standing for node-based framework) are executed as follows:

```
> nodeSim("C01.252.400","C01.539.757",method="SP")
[1] 0.8

> headingSim("Hip","Hand",method="WL",frame="node")
[1] 0.763113


> sa<-c("Body Regions", "Abdomen",
    "Abdominal Cavity")
> sb<-c("Lumbosacral Region", "Body Regions")
> headingSetSim(sa, sb, method="JC", frame="node")
[1] 0.6453512
```

docSim shows the similarity between two documents (PMID).

```
> docSim("2189633", "18974831",frame="heading")
[1] 0.1
```

The next two functions compute the similarities of all pairs of multiple inputs at once. Examples are as follows:

```
> la<-c("B03.440.450.425.800.200",
    "B03.440.450.900.859.225",
    "B01.650.940.800.575",
    "C19.642.355.480.500")
> lb<-c("B03.440.400.425.340",
    "B03.440.400.425.117.800",
    "B03.440.400.425.127.100")
> mnodeSim(la,lb,method="Lord")
          [,1]      [,2]      [,3]
[1,] 0.9962991 0.9962991 0.9962991
[2,] 0.9962991 0.9962991 0.9962991
[3,] 0.8354435 0.8354435 0.8354435
[4,] 0.0000000 0.0000000 0.0000000


> la<-c("Body Regions", "Abdomen",
    "Abdominal Cavity")
> lb<-c("Lumbosacral Region", "Body Regions")
> mheadingSim(la, lb, method="Resnik",
    frame="node")
          [,1]      [,2]
[1,] 0.2967087 0.2967087
[2,] 0.3772228 0.2967087
[3,] 0.3772228 0.2967087
```

The rest three functions show information on queries. *nodeInfo* and *termInfo* are to query MeSH tree information of a given MeSH node and MeSH heading, respectively. The default setting of "brief" is "TRUE", which is to retrieve the whole MeSH tree information including the path to the root node and all child nodes of the given MeSH node or heading. *docInfo* outputs the title, abstract and MeSH headings of a query document, while the default setting is to output all MeSH

heading without the title and abstract (The title and abstract can be printed out by setting "verbose" at "TRUE"). Also "major" can be set at "TRUE" to output major MeSH headings only.

```
> nodeInfo("B03.440.400.425", brief=TRUE)
$B03
[1] "Bacteria"
$B03.440
[1] "Gram-Negative Bacteria"
$B03.440.400
[1] "Gram-Negative Aerobic Bacteria"
$B03.440.400.425
[1] "Gram-Negative Aerobic Rods and Cocci"

> termInfo("United States", brief=TRUE)
[[1]]
[[1]]$Z01
[1] "Geographic Locations"
[[1]]$Z01.107
[1] "Americas"
[[1]]$Z01.107.567
[1] "North America"
[[1]]$Z01.107.567.875
[1] "United States"

> docInfo("1111111", verbose=TRUE, major=TRUE)
[1] "Title: Evaporative water loss in box turtles: effects of rostral
    brainstem and other temperatures."
[1] "Abstract: Box turtles were implanted with thermodes astride the
    preoptic tissue of the brainstem. The rate of evaporative water
    loss could be transiently increased by heating the rostral
    brainstem. Heating tissue in the anterior hypothalamus affected
    evaporative water loss only a high ambient temperatures. The
    magnitude of the response was proportional both to the change in
    hypothalamic temperature and to the ambient temperature with which
     the turtle was in equilibrium. The major fuction of a high rate
    of evaporative water loss in turtles is probably to protect the
    brain from overheating during thermal stress."
[1] "MeSH Headings:"
[1] "Brain Stem"          "Hot Temperature"
[3] "Turtles"             "Water Loss, Insensible"
```

4. Case Studies

4.1. *Application on Biomedical Document Clustering*

As each MEDLINE document corresponds to an MeSH heading set, it has been demonstrated that MeSH plays a critical role in MEDLINE document clustering as one of the most informative field [27,10]. Here, by integrating MeSH semantic similarity into MEDLINE documents clustering, we demonstrate a typical application of the MeSHSim. We choose two popular methods to incorporate MeSH semantic similarity. One of them is proposed by Zhu *et al.* [28] using linear combination (LCM), and the other one, called Semi-supervised Normalized Cut (SSNCut), is introduced by Gu *et al.* [7] which is based on semi-supervised learning.

In the LCM, we first normalize content similarity matrix S^{con} of MEDLINE documents, and semantic similarity matrix S^{sem} of MeSH headings of articles calculated by function *headingSetSim* or *docSim* in the MeSHSim package. Besides, the contents (title and abstract) of documents can be retrieved by function *docInfo*. For simplicity, we denote the normalized S^{con} and S^{sem} as S^{con}_{nor} and S^{sem}_{nor}, respectively. Then, we combine the similarity matrixes linearly by using weight as follows,

$$S_{LCM} = (1 - \omega)S^{con}_{nor} + \omega S^{sem}_{nor} \tag{12}$$

where S_{LCM} is the integrated similarity matrix. Finally, we cluster documents over the S_{LCM}, where Normalized Cut (NCut) is used.

In the SSNCut, it has been demonstrated that MeSH semantic similarity is able to improve performance of MEDLINE documents clustering using semi-supervised learning. Semi-supervised clustering can take advantage of a small amount of prior knowledge to guide clustering process and boost clustering performance. Usually, the prior knowledge is provided as a form of pairwise constraints including must-link and cannot-link. Must-link constraint specifies that a pair of samples must be in the same cluster, and cannot-link constraints suggests a pair of data points cannot be in the same cluster.

SSNCut consists of the following four steps.

(1) Using function *docSim* or *headingSetSim* to calculate the MeSH semantic similarity $Sim^{sem}(d_i, d_j)$ for all pairs of documents (i.e. $d_i \in D$, $d_j \in D$) in document data set. For simplicity, denote the semantic similarity matrix as S^{sem}, where $S^{sem}(i, j)$ is the MeSH semantic similarity between document d_i and d_j.
(2) Using cut-off trick on Sim^{sem} to generate prior constraints for semi-supervised clustering algorithm. For simplicity, we denote ML as must-link set and CL as cannot-link set.
(3) Calculate content similarity $Sim^{con}(d_i, d_j)$ for all pairs of documents (i.e. $d_i \in D$, $d_j \in D$) in document data set. For simplicity, denote the content similarity matrix as S^{con}, where $S^{con}(i, j)$ is the content similarity between document d_i and d_j.
(4) Perform semi-supervised clustering using SSCNut over S^{con}, ML and CL.

4.1.1. *Data set*

In order to demonstrate the usefulness of the MeSHSim, we collected and generated a MEDLINE document data set TREC2005. It is extracted from TREC genomics track 2005[a]. The organisation provides 50 topics to query relevant documents from the TREC Genomics 2005 corpus containing 4,591,008 documents [8]. In our experiments, we regard these 50 topics as true clusters of relevant documents to simulate real information needs in the biomedical domain. To generate our data set, we first

[a]http://ir.ohsu.edu/genomics

remove the topics having only nine or fewer documents, to avoid very small clusters, and further remove documents that are relevant to more than one topic. We then obtain a basic data set of 2,317 documents in 24 topics.

4.1.2. *Evaluation criteria*

To evaluate clustering performance, we choose the Normalized Mutual Information (NMI) which is a popular and well-accepted criteria in clustering domain. It has been proven that NMI show better performance than other criteria [6]. Since there are multiple version of NMI, we use the squared version proposed by [23] defined as follows,

$$NMI(P;C) = \frac{I(P;C)}{\sqrt{H(P) * H(C)}} \tag{13}$$

where P and C are the predicted labels and true labels, respectively. $I(P;C) = H(P) - H(P|C)$ is the mutual information between P and C. $H(.)$ stands for entropy.

4.1.3. *Results*

Table 2 reports the clustering results using the LCM. The first column shows result using NCut over S_{nor}^{con} without MeSH similarity. The other four columns presents results based on S_{LCM}, and the numbers in the titles imply the weights ω of MeSH semantic similarity S_{nor}^{sem}.

It is obvious that the LCM outperforms the NCut using only content similarity. For example, the NMI of LCM is improved to 0.8210 at least, and 0.8518 at most, while the NMI of NCut is only 0.8022. It is demonstrated that our MeSHSim package is able to help researchers analyze MELDINE documents.

Table 2. NMI results of LCM on data set TREC2005.

NCut	$LCM_{0.3}$	$LCM_{0.4}$	$LCM_{0.5}$	$LCM_{0.6}$
0.8022	0.8302	**0.8518**	0.8286	0.8210

Table 3. NMI results of SSNCut on data set TREC2005.

NCut	$SSNCut_{0.5\%}$	$SSNCut_{1\%}$	$SSNCut_{2\%}$	$SSNCut_{5\%}$
0.8022	0.8719	**0.8770**	0.8756	0.8671

Table 3 shows the clustering results using the SSNCut. The last four columns list results combining MeSH similarity, which is used to generate constraints. The percentage numbers are the ratio of generated constraints using cut-off trick. For example, $SSNCut_1\%$ means the pairs of documents, whose MeSH similarities are within top 1%, are connected with must-links, while those within bottom 1% are used to generate cannot-links. Note that we use the "JC" method and "node" based framework in the function *headingSetSim* to calculate MeSH-based similarities over MEDLINE articles.

As shown in the Table 3, the SSNCut outperforms NCut at all the four percentages of constraints. For instance, the NMI of SSNCut is boosted to 0.8671 at least, and 0.8770 at most, while the NMI of NCut is only 0.8022. Thus, we can see that our MeSHSim package is able to provide convenience for researchers to do MEDLINE documents clustering.

4.2. *Application on Searching for Similar Concepts*

Searching for similar concepts plays a critical roles in medical information retrieval. For example, expanding users' queries with similar MeSH concepts should help promote a better understanding of users' intention and improve the performance of retrieving [117]. Here, we pick eight MeSH heading randomly, and use function *headingSim* to search for the most similar MeSH concepts. Table 4 gives searching results the using different measures with heading-based framework, while Table 5 shows results with node-based framework. The numbers under the similar MeSH headings are the similarities between related MeSH headings.

As shown in Table 4 and Table 5, the results of path-based methods differ from the ones of IC-based methods typically. For example, in the case of the MeSH heading "Tribolium", since "Beetles" (B01.050.500.131.617.069) is the parent node of the "Tribolium" (B01.050.500.131.617.069.799), the three path-based measures believe it is the most similar concept of the "Tribolium". "Weevils", however, is the most similar concept to "Tribolium" assessed by the method Lord, which only takes the nearest common ancestor into account. This is because "Weevils" (B01.050.500.131.617.069.900) and "Tribolium" (B01.050.500.131.617.069.799) are brothers whose parent is "Beetles", and the deeper the heading is, the more information it has. As for the method Lin and JC, "Ephemeroptera" (B01.050.500.131.617.104) is considered as the most similar concept to "Tribolium" after taking the information of headings themselves into account besides the nearest common ancestor.

Besides, we can see that the results of MeSH headings "Ovomucin", "Glycosuria", "Chromates", "Archives" and "Chest Pain" under heading-based framework is different from the ones under node-based framework. That is because these five MeSH headings are presented by several MeSH nodes. For instance, we can see that "Ovomucin" corresponds to MeSH nodes "12.776.256.317.675", "D12.776.395.560.760" and "D12.776.290.675" by using function *termInfo* in our MeSHSim package. Therefore, the MeSHSim can promote better understanding of MeSH concepts and finds similar MeSH headings to help other biomedical text analysis tasks.

5. Conclusion

The measures of the semantic similarities for MeSH ontology facilitate users to compare MEDLINE documents, and therefore have became an significant prior knowledge in many text mining approaches in biomedical domain. The MeSHSim package

Table 4. Application of Searching for the Most Similar MeSH Concepts using MeSHSim with heading-based framework

Heading	SP	WP	Li	Lord	Lin	JC
Ovomucin	Egg Proteins 0.8864	Egg Proteins 0.7444	Egg Proteins 0.6154	Avidin 0.9845	Avidin 0.5917	Egg Proteins 0.1402
Glycosuria	Urination Disorders 0.8900	Urination Disorders 0.7636	Urination Disorders 0.6745	Nephrocalcinosis 0.9953	Urination Disorders 0.6968	Pregnancy, Interstitial 0.3202
Chromates	Anions 0.8939	Anions 0.7542	Anions 0.6453	Chromium Compounds 0.9872	Chromium Compounds 0.7366	Chromium Compounds 0.4834
Tribolium	Beetles 0.9524	Beetles 0.9476	Beetles 0.8183	Weevils 0.9999	Ephemeroptera 1.0894	Ephemeroptera 1.4474
Catheters	Urinary Catheters 0.9474	Urinary Catheters 0.8821	Urinary Catheters 0.7752	Urinary Catheters 0.9999	Catheters, Indwelling 0.9898	Catheters, Indwelling 0.9379
Amoeba	Tubulina 0.9524	Tubulina 0.9231	Tubulina 0.8175	Endolimax 0.9999	Tubulina 0.9438	Tubulina 0.6716
Archives	Knowledge 0.7636	Museums 0.5926	Museums 0.5369	Knowledge 0.9836	Museums 0.6553	Sverdlovsk Accidental Release 0.7057
Chest Pain	Flank Pain 0.9000	Flank Pain 0.8000	Flank Pain 0.6594	Flank Pain 0.9984	Signs and Symptoms 0.7166	Allesthesia 0.7305

Table 5. Application of Searching for the Most Similar MeSH Concepts using MeSHSim with node-based framework

Heading	SP	WP	Li	Lord	Lin	JC
Ovomucin	Mucoproteins 0.9545	Mucoproteins 0.9091	Mucoproteins 0.8147	Egg Proteins, Dietary 0.9999	Egg Proteins, Dietary 0.8391	Egg Proteins, Dietary 0.3063
Glycosuria	Glucose Metabolism Disorders 0.9500	Glycosuria, Renal 0.9231	Glycosuria, Renal 0.8175	Glycosuria, Renal 0.9999	Pregnancy, Interstitial 0.8896	Pregnancy, Interstitial 0.6735
Chromates	Potassium Dichromate 0.9545	Anions 0.9091	Anions 0.8147	Potassium Dichromate 0.9999	Chromium Compounds 0.9543	Chromium Compounds 0.7219
Tribolium	Beetles 0.9524	Beetles 0.9333	Beetles 0.8183	Weevils 0.9999	Ephemeroptera 1.0894	Ephemeroptera 1.4474
Catheters	Urinary Catheters 0.9474	Urinary Catheters 0.8571	Urinary Catheters 0.7752	Urinary Catheters 0.9999	Catheters, Indwelling 0.9898	Catheters, Indwelling 0.9379
Amoeba	Tubulina 0.9524	Tubulina 0.9231	Tubulina 0.8175	Endolimax 0.9999	Tubulina 0.9438	Tubulina 0.6716
Archives	Information Centers 0.9333	Museums 0.8888	Museums 0.8054	Museums 0.9999	Sverdlovsk Accidental Release 1.0950	Sverdlovsk Accidental Release 1.4114
Chest Pain	Angina Pectoris 0.9500	Angina Pectoris 0.9091	Angina Pectoris 0.8147	Angina Pectoris 0.9997	Allesthesia 1.1455	Allesthesia 1.4875

implemented nine typical MeSH ontology-based semantic similarity measures in the powerful R system. Compared with few existing related tools, like the online server THE MESH SIMILARITY, the MeSHSim can be easily integrated into pipelines for other biomedical text analysis task to improve their performance, such as information retrieval, biomedical document clustering and citation searching process. Other

utilities, such as functions for querying MeSH heading information and retrieving MEDLINE documents, should offer a straightforward way to study MeSH tree and MEDLINE documents. In all, we provide a general software to calculate semantic similarity over MeSH headings and MEDLINE documents. Related researchers are able to develop specific domain applications based on MeSHSim according to their needs, for example, calculating similarity over diseases, drugs, cures and biomedical documents, etc.

Acknowledgments

This work has been partially supported by National Natural Science Foundation of China (61572139, 61272110), Scientific Research Starting Foundation for Returned Overseas Chinese Scholars, Ministry of Education, China and JSPS KAKENHI (#2430054).

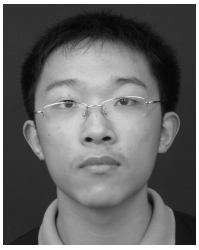
References

1. Azcárate MC, Vázquez JM, López MM, Improving image retrieval effectiveness via query expansion using mesh hierarchical structure, *Journal of the American Medical Informatics Association* **20**(6):1014–1020, 2013.
2. Bien SJ, Park CH, Shim HJ, Yang W, Kim J, Kim JH, Bi-directional semantic similarity for gene ontology to optimize biological and clinical analyses, *JAMIA* **19**(5):765–774, 2012.
3. Blott S, Gurrin C, Jones G, Smeaton A, Sodring T, On the use of mesh headings to improve retrieval effectiveness., *Text REtrieval Conference (TREC2003)* pp. 215–224, 2003.
4. Bulskov H, Knappe R, Andreasen T, On measuring similarity for conceptual querying., *Proceedings of the 5th International Conference on Flexible Query Answering Systems (FQAS)* **2522**:100–111, 2002.
5. Cohen PR, Kjeldsen R, Information retrieval by constrained spreading activation in semantic networks, *Information processing & management* **23**(4):255–268, 1987.
6. Ghosh J, Scalable clustering methods for data mining, *Handbook of data mining, N Ye, Ed Lawrence Erlbaum* , 2003.
7. Gu J, Feng W, Zeng J, Mamitsuka H, Zhu S, Efficient semisupervised medline document clustering with mesh-semantic and global-content constraints, *IEEE Trans Cybernetics* pp. 1265–1276, 2013.
8. Hersh W, Cohen A, Yang J, Bhupatiraju RT, Roberts P, Hearst M, Trec 2005 genomics track overview, *In TREC 2005 notebook*, pp. 14–25, 2005.
9. Hliaoutakis A, Varelas G, Voutsakis E, Petrakis EG, Milios E, Information retrieval by semantic similarity, *International journal on semantic Web and information systems (IJSWIS)* **2**(3):55–73, 2006.
10. Huang X, Zheng X, Yuan W, Wang F, Zhu S, Enhanced clustering of biomedical documents using ensemble non-negative matrix factorization, *Information Sciences* **181**(11):2293–2302, 2011.
11. Jiang J, Conrath D, Semantic similarity based on corpus statistics and lexical taxonomy., *In Proceedings of the International Conference on Research in Computational Linguistic, Taiwan* , 1998.
12. Lang D, R as a web client-the rcurl package., *Journal of Statistical Software* , 2007.
13. Leacock C, Chodorow M, Filling in a sparse training space forword sense identification., *In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics(ACL94)* , 1994.

14. Li Y, Bandar ZA, McLean D, An approach for measuring semantic similarity between words using multiple information sources., *IEEE Transactions on Knowledge and Data Engineering* **15**(4):871–882, 2003.
15. Lin D, Principle-based parsing without overgeneration, *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*ACL '93, ACL '93, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 112–120, 1993. doi:10.3115/981574.981590, URL `http://dx.doi.org/10.3115/981574.981590`.
16. Lord P, Stevens R, Brass A, Goble C, Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation., *Bioinformatics* **19**(10):1275–1283, 2003.
17. Lu Z, Kim W, Wilbur WJ, Evaluation of query expansion using mesh in pubmed, *Information retrieval* **12**(1):69–80, 2009.
18. Nelson S, Schopen M, Savage A, Schulman J, Arluk N, The mesh translation maintenance system: structure, interface design, and implementation., *Studies in health technology and informatics* **107**(Pt 1):67–69, 2004.
19. Névéol A, Zeng K, Bodenreider O, Besides precision & recall: Exploring alternative approaches to evaluating an automatic indexing tool for medline, *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, p. 589, 2006.
20. Rada R, Mili H, Bichnell E, Blettner M, Development and application of a metric on semantic nets., *IEEE Trans Syst, Man, Cybern* **9**:17–30, 1989.
21. Resnik O, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity and natural language., *Journal of Artificial Intelligence Research* **19**:95–1130, 1999.
22. Richardson R, Smeaton A, Murphy J, *Using WordNet as a knowledge base for measuring semantic similarity between words*, 1994.
23. Strehl A, Ghosh J, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, *The Journal of Machine Learning Research* **3**:583–617, 2003.
24. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF, A new method to measure the semantic similarity of go terms, *Bioinformatics* **23**(10):1274–1281, 2007. doi:10.1093/bioinformatics/btm087, URL `http://bioinformatics.oxfordjournals.org/content/23/10/1274.abstract`.
25. Wu Z, Palmer M, Verbs semantics and lexical selection., *In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics (ACL'94)* pp. 133–138, 1994.
26. Yoo I, Hu X, Song IY, Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 791–796, 2006.
27. Zhu S, Takigawa I, Zeng J, Mamitsuka H, Field independent probabilistic model for clustering multi-field documents, *Information Processing & Management* **45**(5):555–570, 2009.
28. Zhu S, Zeng J, Mamitsuka H, Enhancing medline document clustering by incorporating mesh semantic similarity., *Bioinformatics* **25**(15):1944–1951, 2009.

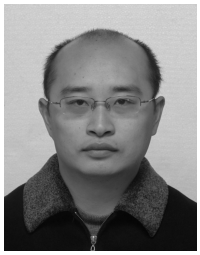
Jing Zhou received his B.S. degree in computer science from Donghua University, Shanghai, in 2012.

He is currently a graduate student of the Shanghai Key Laboratory of Intelligent Information Processing, Fudan University. His current research interests include data mining, bioinformatics and time series analysis.



Yuxuan Shui received his B.S. degree in computer science from Fudan University, Shanghai. Currently a graduate student in computer science at Stony Brook University.

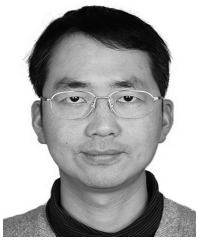
Shengwen Peng received his B.S. degree in computer science from Chang'an University, Xian, in 2014. He is currently a master student of the Shanghai Key Laboratory of Intelligent Information Processing, Fudan University. His current research interests include multi-label classification, data mining, and bioinformatics.



Xuhui Li is an Associate Professor at Wuhan University, China. He received his Ph.D in Computer Science from Wuhan University in 2003. His research interests include data semantics, linguistic semantics, knowledge engineering and formal methods.

Hiroshi Mamitsuka received the B.S. degree in biophysics and biochemistry, the M.E. degree in information engineering, and the Ph.D. degree in information sciences from the University of Tokyo, Tokyo, Japan, in 1988, 1991, and 1999, respectively.

He is currently a Professor with Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. His current research interests include mining from graphs and networks in biology and chemistry.



Shanfeng Zhu Shanfeng Zhu received the B.S. and M.Phil. degrees in computer science from Wuhan University,Wuhan, China, in 1996 and 1999, respectively, and the Ph.D. in computer science from the City University of Hong Kong, Hong Kong, in 2003. He was a Postdoctoral Fellow with the Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. Since July 2008, he has been with the School of Computer Science and Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China, where he is currently an Associate Professor. His research focuses on developing and applying machine learning, data mining, and algorithmic methods for bioinformatics and information retrieval. Dr. Zhu is a member of China Computer Federation and the Association for Computing Machinery.