Towards Transforming Expert-based Content to Evidence-based Content

Soheil Moosavinasab, BS[1, 2]; Majid Rastegar-Mojarad, MS[1, 2]; Hongfang Liu, PhD[1]; Siddhartha R. Jonnalagadda, PhD[1, 3]

[1] Department of Health Sciences Research, Mayo Clinic, Rochester, MN

[2] University of Wisconsin-Milwaukee, Milwaukee, WI

[3] Northwestern University Feinberg School of Medicine, Chicago, IL

Abstract

The goal of this paper is to find relevant citations for clinicians' written content and make it more reliable by adding scientific articles as references and enabling the clinicians to easily update it using new information. The proposed approach uses information retrieval and ranking techniques to extract and rank relevant citations from MEDLINE for any given sentence. Additionally, this system extracts snippets of relevant content from ranked citations. We assessed our approach on 4,697 MEDLINE papers and their corresponding full-text on the subject of Heart Failure. We implemented multi-level and weight ranking algorithms to rank the citations. We demonstrate that using journal relevance and study design type improves results obtained from only using content similarity by approximately 40%. We also show that using full-text, rather than abstract text, leads to extracting higher quality snippets.

Introduction

In this paper, we developed a system, known as CiteFinder, to find citations for clinical sentences. For each given sentence, the system finds citations from MEDLINE articles, ranks the citations based on similarity with the sentence, and extracts a snippet for each citation. We implemented a tool for the system that allows the user to submit a sentence and receive back the top relevant citations. This aids in transforming the expert-based content (a paradigm not used by certain clinical knowledge systems such as UpToDate©[1], but relatively common among some care providers[2]) to evidence-based content – the accepted paradigm[3]. This will offer clinicians the flexibility of easily authoring evidence-based guidance and FAQs for their peers.

Background

Citation finding has been investigated to recommend relevant papers to researchers[4–8]. There are also studies on information retrieval in the medical domain. For example, Plaza and Diaz[9] proposed a method to query similar Electronic Health Records using UMLS concepts. Hersh and Hickam[10] studied the effectiveness of electronic information retrieval systems for physicians. Lu[11] investigated web tools for searches in biomedical literature. Bachmann et al [12] proposed and validated search strategies used to identify diagnostic articles recorded on MEDLINE, with special emphasis on precision. Bernstam et al[13] studied how citation-based algorithms that are developed to extract relevant and important citations for the World Wide Web are useful in the biomedical literature domain. They compared eight citation algorithms, including simple PubMed queries, clinical queries, citation counts, journal impact factors, etc. Their research concluded that these citation-based algorithms are useful in the domain of biomedical literature. Lin et al[14] extracted relevant MEDLINE citations and ranked them based on several ranking methods, including citation counts per year and journal impact factors. Darmoni et al[15] used MeSH concepts for indexing and information retrieval. Some studies have also been conducted on query expansion using MeSH terms in PubMed. Lu et al[16] analyzed the effect of using MeSH terms in a PubMed automatic search. In the current study, we also used MeSH concepts to find relevant citations.

Methods

CiteFinder consists of four major parts: sentence expansion, citation extraction, citation ranking, and snippet generation.

After a user submits a sentence (although technically this could be applied for an entire paragraph), the system finds relevant citations for the sentence from our collection of MEDLINE articles. To find relevant citations, MeSH terms are used. CiteFinder extracts MeSH terms from the sentence and searches them in MeSH terms of each indexed MEDLINE article. Then it ranks the articles based on three measures: MeSH terms, journal relevance, and epidemiological study design[17]. The final step is producing snippets for the retrieved citations based on the extracted major terms (mentions) of the sentence.

Figure 1 illustrates the architecture of the system. We use a running example in Appendix 1 to clarify each part of the system.

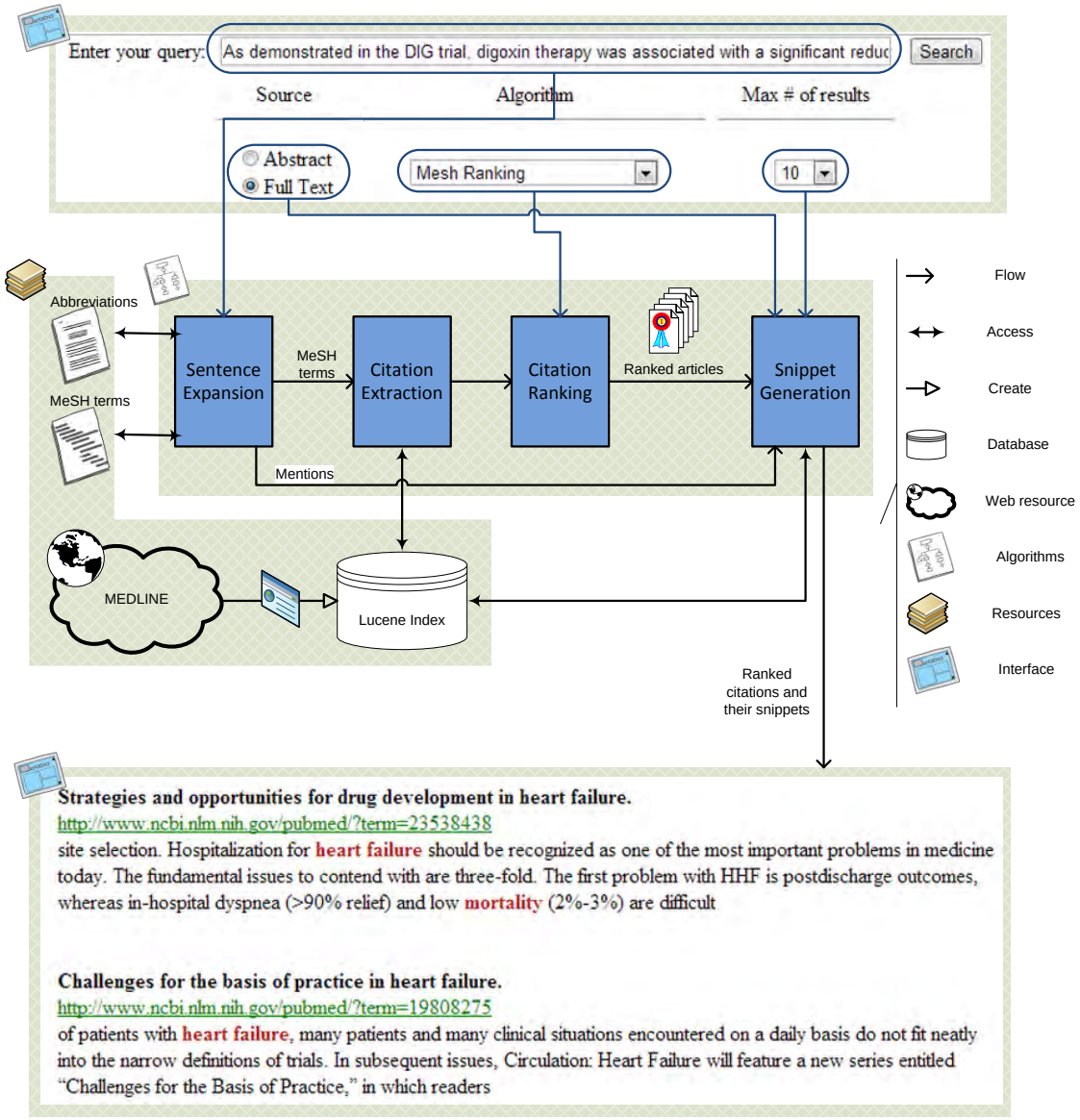

Figure 1: System Architecture The figure illustrates the sentence expansion, citation extraction, citation ranking, and snippet generation components and their integration with the user-interface – all of them available at http://sourceforge.net/projects/cksauthorer.

Step 1: Sentence Expansion Since each word in a sentence might not be in an article or abstract, we locate important terms, normalize them and expand. That is, the sentence goes through OpenNLP tokenization[18], lexical normalization[19], dictionary-based concept extraction using both UMLS Metathesaurus and MeSH using Aho-Corasick algorithm[20], and abbreviation expansion (using a list of 6,024 abbreviations and their full-forms derived from UMLS).

Step 2: Citation Extraction The next step is to find relevant citations for the sentence based on the extracted MeSH terms. To be able to generalize the system to other documents such as textbooks and guidelines and build a fast system, we indexed MEDLINE abstracts and their full-text with Lucene[21]. CiteFinder stores the text, title,

publication type, and MeSH terms of each article. The articles with at least one MeSH term in common with the sentence will be retrieved at this step.

Step 3: Citation Ranking In order to rank the retrieved citations with regard to their importance and similarity with the sentence, three measures are applied: MeSH ranking, journal relevance, and study design. In the following section, we describe each of them and explain how we calculate a score.

Measure 1: MeSH Measure. The MeSH measure shows the semantic similarity of the sentence and articles. We use the score calculated by Lucene for each returned article from the MeSH extraction step. Our language model that is based on Mesh terms as opposed to individual words built from Lucene[22] takes into account both the TF-IDF (frequency of the term in the document with penalty to each term if it is commonly occurring in other documents), and Number of MeSH terms in an article. This performed better than TF-IDF over individual words.

Measure 2: Journal Relevance. The idea behind this measure is that a citation that is published in a high-quality journal has extra chance to obtain a higher rank than a citation with the same MeSH score published in a low-quality journal in specific domain (example, Heart Failure [CHF]). We previously studied the task of prioritizing journals and obtained a formula to rank each journal[23] based on information available from Scopus[24] and PubMed – Journal Relevance score = (0.82640 * SCImago Journal Ranking) – (0.00377 * Number of articles) + (0.00258 * Number of articles for 3 years) – (0.00190 * Number of cited-articles for 3 years) – (0.01846 * Number of references per article) + (0.00295 * Number of CHF-indexed Medline abstracts) + (0.62864 * Is Broad Journal Heading cardiology?) – (0.32753 * Is Core clinical journal?).

Measure 3: Study Design. It is well known that the strength of the findings in clinical research depends on the study design and follows this order: systematic review, randomized controlled trial, multiple time series, nonrandomized trial, cohort, case-control, time series, cross-sectional, and case series[17]. Weight levels 9 to 1 are assigned to each study type, respectively. To decide on the study type of a citation, we consider several sections of articles, including publication type, abstract text, MeSH headings, and article title. A publication for which no study design is detected gets the least possible score of 1.

Ranking Methods

We proposed two ranking schemes using the above measures to assign ranks to retrieved citations. It should be noted that all scores of the measures are normalized to the range of 0 to 1.

1) ***Multi-Level Ranking.*** A multi-level approach ranks the articles in a cascade trend. The idea is to rank the articles with one measure, and then split the sorted articles into brackets and re-rank the brackets with scores obtained from other measures. Finding the best bracket size for each level is one of the challenges of this approach. In this experiment, after extracting and ranking citations via MeSH measure, CiteFinder splits them in N brackets based on their MeSH score. Table 1 shows the results with different variation of N. In the next step, the journal measure is used to rank the citations within the bracket. In the last step, the study design measure is used to rank the citations in each newly created N brackets to produce the final list of ranked citations.

2) ***Weight Ranking.*** In the second approach, the final score is calculated using the formula: Score = (MeSH weight * MeSH score) + (Journal Relevance weight * Journal Relevance score) + (Study Design weight * Study Design score). This approach is valid considering that these three metrics are independent and orthogonal.

Step 4: Snippet Generation. Snippet generation is helpful for clinicians to get an idea about the existence of similar information in scholarly articles and improve their written content. A query made by disjunction of the extracted mentions is used to extract a maximum of three snippets for each article. See Appendix 1 for a running example of all the steps.

Evaluation

Data Collection. CiteFinder contains 4,697 MEDLINE papers about Heart Failure. This corpus includes two major sources (the duplicated articles or the articles with only scanned-version availability have been removed):

- 2,582 articles retrieved by “heart failure[MeSH Major Topic]” query at PubMed Central

- 2,262 articles retrieved by “Congestive Heart Failure[MeSH Major Topic]” query at PubMed on four top ranked journals for CHF topic: 1. Circulation, 2. Circulation. Heart failure, 3. JAMA the Journal of the American Medical Association, and 4. The New England Journal of Medicine

Both the abstract and full-text of these papers are indexed separately with Lucene to allow us to compare the performance on both the abstract and full-text in extracting snippets.

Gold standard

The gold standard data contains 377 sentences referring to 456 citations. We primarily selected 7,864 sentences referring to 11,778 citations using all 150 retrieved articles from UpToDate© for the query – "heart failure". We then filtered out sentences with less than 15 words, less than 5 MeSH terms, or no full-text availability in our index files.

Results

To evaluate ranking methods, we consider median rank of expected citations for each sentence in our gold standard. If the expected citation of a sentence is not retrieved, its rank is assumed to be the worst (lowest). So we consider the median rank of all citations in the gold standard, regardless of whether the system finds and ranks them or not. In this scenario, we were unable to find 5.26% (24 of 456) of the citations, but the currently reported median ranking is affected by recall.

Multi-Level Ranking

In the first experiment, we explored multi-level ranking. Table 1 shows median rank for the multi-level ranking approach. Both Journal Relevance and Study Design show improvement in the results.

Table 1: Multi-level ranking results

Measures \ # of Brackets	10	20	100
MeSH	76	76	76
MeSH and Journal Relevance	66	67	65
MeSH and Study Design	73	71	67
MeSH, Journal Relevance, and Study Design	**64**	**65**	**65**

Weight Ranking

In the second experiment, we attempted to find the best coefficient for Journal Relevance when the MeSH coefficient is 1. A range of coefficients between 0 and 2 were explored, and the results indicated 0.5 as the best weight for Journal Relevance Figure 2 illustrates these results.

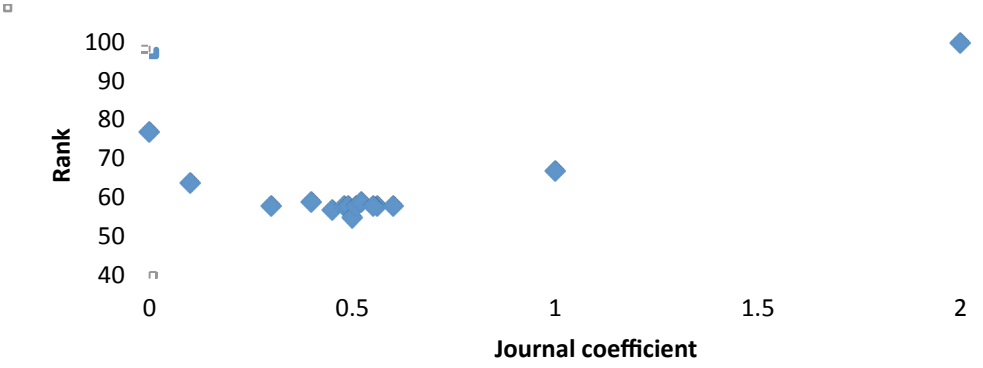

Figure 2: The chart indicates how different journals coefficients affect median ranking in our gold standard. In this experiment, the coefficient for MeSH measure is 1.

Then we used the best combination (MeSH=1, Journal Relevance=0.5) as the constant and found the best weight for Study Design (0.30). Figure 3 indicates the results of this experiment.

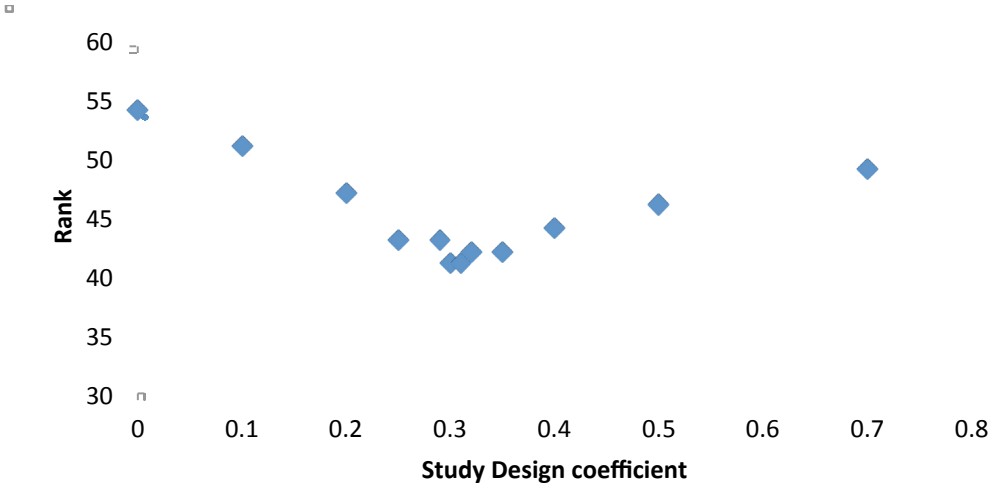

Figure 3: The chart indicates how different Study Design coefficients affect median ranking in our gold standard. In this experiment, the coefficient for MeSH measure is 1 and Journal Relevance is 0.5.

Snippet Generation After ranking returned citations, we extracted snippets for each of them. In this experiment, we explored whether using the full-text for extracting snippets is better than using the abstract. The experiment on the gold standard indicated that when CiteFinder uses full-text, it is able to extract at least one snippet for 99.7% of citations (in 431 of 432 extracted and ranked citations). When the system looks for snippets in an abstract, it extracted snippets for 80.7% of the citations (349 out of 432). Further, as the system tries to extract the best snippet (as adjudicated by the MeSH-based Lucene similarity), we discovered that only 22.58% of the best snippets come from the abstract text with the rest coming from full-text. This means that using the full-text instead of abstract text leads to the collection of more and better snippets.

Discussion

Ranking Algorithm. We implemented both multi-level and weight ranking algorithms to rank the citations. Results show more improvement in the weight-ranking algorithm because of the flexibility of this approach to change the effectiveness of measures. On the other hand, the multi-level approach is sensitive to the number of results retrieved by CiteFinder. In cases where the number of retrieved articles is not considerably larger than number of brackets, the system will not actually utilize the second- or third-level measures.

Generalizability. The proposed system (CiteFinder) explores methods to find citations for sentences in the Heart Failure domain. Further experiments will be required to check the generalizability of the system in other domains. Future work should also explore better methods to infer the epidemiological study design of the publication and consider alternative ways to score them. Appendix 2 discusses further limitations.

Conclusions

Finding supporting citations for clinical sentences is challenging for clinicians. We propose a system (CiteFinder), which, after expanding a user's sentence, extracts relevant citations and ranks them to retrieve the best citation for a given sentence. This study demonstrates that using Journal Relevance and Study Design type will improve the MeSH term results by about 40% (from 76 to 41). We also show that using full-text instead of abstract-text helps in extracting better snippets; i.e., they have more pertinent information corresponding to the input queries. The code for various components including the user-interface is available at http://sourceforge.net/projects/cksauthorer.

Acknowledgments

This work was made possible by joint funding from the National Library of Medicine K99/R00 LM011389, National Institute of Health R01LM009959A1, and National Science Foundation ABI:0845523. We are also thankful to UpToDate© for allowing us to use their data for research purposes.

References

1. UpToDate [Internet]. Available from: http://www.uptodate.com/home

2. Yeo GSH, Lim ML. Maternal and fetal best interests in day-to-day obstetrics. Ann Acad Med Singap 2011;40(1):43–9.

3. Lau J. Evidence-based medicine and meta-analysis: getting more out of the literature. Clinical decision support: the road ahead 2007;

4. Huang Z, Chung W, Ong T, Chen H. A graph-based recommender system for digital library. In: In Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press; 2002. p. 65–73.

5. Chen Y-L, Wei J-J, Wu S-Y, Hu Y-H. A similarity-based method for retrieving documents from the SCI/SSCI database. Journal of Information Science 2006;32(5):449–64.

6. Ratprasartporn N, Po J, Cakmak A, Bani-Ahmad S, Ozsoyoglu G. Context-based literature digital collection search. The VLDB Journal 2009;18(1):277–301.

7. Bollacker KD, Lawrence S, Giles CL. Discovering Relevant Scientific Literature on the Web. IEEE Intelligent Systems 2000;15(2):42–7.

8. Liang Y, Li Q, Qian T. Finding Relevant Papers Based on Citation Relations [Internet]. In: Wang H, Li S, Oyama S, Hu X, Qian T, editors. Web-Age Information Management. Springer Berlin Heidelberg; 2011 [cited 2013 Aug 14]. p. 403–14.Available from: http://link.springer.com/chapter/10.1007/978-3-642-23535-1_35

9. Plaza L, Díaz A. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs [Internet]. In: Hopfe CJ, Rezgui Y, Métais E, Preece A, Li H, editors. Natural Language Processing and Information Systems. Springer Berlin Heidelberg; 2010 [cited 2013 Aug 14]. p. 296–303.Available from: http://link.springer.com/chapter/10.1007/978-3-642-13881-2_31

10. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998;280(15):1347–52.

11. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford) 2011;2011:baq036.

12. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc 2002;9(6):653–8.

13. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using citation data to improve retrieval from MEDLINE. J Am Med Inform Assoc 2006;13(1):96–105.

14. Lin Y, Li W, Chen K, Liu Y. A Document Clustering and Ranking System for Exploring MEDLINE Citations. J Am Med Inform Assoc 2007;14(5):651–61.

15. Darmoni SJ, Soualmia LF, Letord C, et al. Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases. J Med Libr Assoc 2012;100(3):176–83.

16. Lu Z, Kim W, Wilbur WJ. Evaluation of Query Expansion Using MeSH in PubMed. Inf Retr Boston 2009;12(1):69–80.

17. Fletcher RH, Fletcher SW, Fletcher GS. Clinical Epidemiology: The Essentials. Lippincott Williams & Wilkins; 2012.

18. Baldridge J, Morton T. OpenNLP. 2004.

19. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.

20. Aho AV, Corasick MJ. Efficient string matching: an aid to bibliographic search. Commun ACM 1975;18(6):333–40.

21. Lucene [Internet]. Available from: http://lucene.apache.org/

22. Similarity in Lucene [Internet]. Available from: http://lucene.apache.org/core/3_0_3/api/core/org/apache/lucene/search/Similarity.html

23. Jonnalagadda S, Moosavinasab S, Li D, Abel M, Chute C, Liu H. Prioritizing journals relevant to a topic for addressing clinicians' information needs. 2013 (Under review).

24. SCImago. 2007;Available from: Available: http://www.scimagojr.com/

Appendix 1: Running example.

Here we use a running example to demonstrate input, output and results of the system in different steps.

Input query:

For patients who are still hypertensive after initiation of beta blockers and ACE inhibitors and/or ARBs or who cannot tolerate these drugs appropriate agents include loop diuretics nitrates hydralazine and some vasoselective calcium channel blockers (eg amlodipine and felodipine)

Expected citation for this sentence: *PMID9264493*

Abbreviations found:

ACE: angiotensin-converting enzyme

Extracted Mentions:

Hypertensive, initiation, beta blockers, ACE inhibitors, drugs, agents, loop diuretics, nitrates, hydralazine, calcium channel blockers, amlodipine, felodipine, angiotensin receptor blocker, angiotensin, receptor, blocker

Extracted MeSH Terms:

Adrenergic beta-Antagonists, Angiotensin-Converting Enzyme Inhibitors, Pharmaceutical Preparations, Diuretics, Nitrates, Hydralazine, Calcium Channel Blockers, Amlodipine, Felodipine

Rank using the multi-level ranking method: *6th*

Rank using the weight ranking method: *6th*

Query to extract snippet:

""CA" "blocker" "receptor" "angiotensin" "angiotensin receptor blocker" "felodipine" "amlodipine" "calcium channel blockers" "hydralazine" "nitrates" "loop diuretics" "agents" "drugs" "ACE inhibitors" "beta blockers" "initiation" "hypertensive""

The extracted snippet:

antagonists; use of blockers, long-acting nitrates, or other vasodilators (except ACE inhibitors...V-HeFT III Abstract Background Despite therapy with diuretics, ACE inhibitors and digoxin morbidity... or volume, which are reduced by nitrates and ACE inhibitors. Progressive LV remodeling is characterized

Appendix 2: Limitation of Measures.

MeSH Accessibility

MeSH measure is the main method we are using to rank the citations. Journal rank and study design type are added as a component to the MeSH measure to improve the results. All the articles that we have in our corpus are extracted from PubMed or PubMed Central provides MeSH terms for them. CiteFinder's main limitation is that if we want to expand the corpus to cover more articles from mentioned sources, we will need to use a MeSH extractor program to pull out and index the MeSH terms from the articles.

Journal Relevance Measure

We studied 23 sentences related to heart failure with 31 citations. The study shows that 31% of retrieved articles (12,362 of 39,839) were not from the 63 journals we already have. Having a list of important Heart Failure–related journals will automatically guarantee that many unavailable journals are not related to the query. Even though we should assign a score of zero to them, having a complete list of journals can improve the system.

Study Design

We assigned weights of 1 through 9 to different study design types. Machine Learning algorithms can be applied to assign more accurate and meaningful weights to the elements.

Appendix 3: Detailed results for determining the best coefficients

Table 2: Journal coefficient impact on MeSH Ranking (MeSH=1)

Journal Coefficient	0	0.1	0.3	0.4	0.45	0.48	0.49	0.5	0.51	0.52	0.55	0.6	1	2
Median Rank	76	63	57	58	56	57	57	54	57	58	57	57	66	99

Table 3: Study Design coefficient impact on MeSH and Journal Relevance Ranking (MeSH=1, Journal Relevance=0.5)

Study Design Coefficient	0	0.1	0.2	0.25	0.29	0.3	0.31	0.32	0.35	0.4	0.5	0.7
Median Rank	54	51	47	43	43	41	41	42	42	44	46	49