Expanding Queries with Semantically Related Terms Derived from Knowledge Sources

ABSTRACT

Using query expansion to handle the problem of query-document mismatch has been studied for decades. Deriving query expansion terms has been a critical step in the expansion process and has been intensively investigated. Past efforts focused on deriving statistically related terms. In this paper, we propose a methodology that utilizes a domain-specific knowledge source to select semantically related terms. Experimental results reveal that the knowledge-based approach results in a more focused query expansion and is easier for the user to understand. Further, the expanded query is shorter and significantly saves computation time. Finally, the knowledge-query expansion also yields higher precision and recall than the expansion without using any knowledge.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation*

General Terms

Algorithms, Performance, Experimentation

Keywords

Automatic query expansion, knowledge source.

1. INTRODUCTION

Query-document mismatch is a fundamental problem in information retrieval. Queries are usually phrased with limited or general terms, while full-text documents often describe the same topic in different expressions or specific terms. Terms in the original queries often do not match well with the terms in relevant documents. Consequently, relevant documents cannot be retrieved or ranked at the top of the return list, resulting in poor retrieval performance.

Query expansion supplements the original query with additional related terms and increases the chance for the modified query to match with relevant documents. The method has been widely accepted by the literature as an effective means to ameliorate the query-document mismatch problem [3]. Depending on the source where the additional terms originate, query expansion methods can be briefly classified into three categories:

1) Manual query expansion. A human expert manually looks at the original query and picks the most relevant terms to expand. As reported in [17, 18], it is a difficult and tedious task to select the right terms to improve the retrieval result.

2) Relevance feedback. The user needs to mark out relevant documents for the original query, after which terms from relevant documents are used for expansion [10, 12, 14]. A relevance feedback system has to first interact with the user to obtain the relevancy judgements, which, in some cases, creates extra workload for the end user.

3) Automatic query expansion. The retrieval system refers to either a term co-occurrence thesaurus and/or a human generated knowledge source to generate a list of expansion terms, as well as the weight assigned to each term. No interaction with the end user is needed in the expansion process, which is different from 1) and 2).

Specific automatic expansion methods have been proposed in the past three decades. In the pioneering work of [15, 16], Sparck-Jones classifies terms into equivalence classes based on pair-wise term similarity. A query is expanded by all the additional terms in the same classes as the query terms. However, due to the limited size of the sample corpus and potential errors introduced during term classification, little improvement has been reported. [1, 2] employ a similar idea of creating term equivalence classes through document classification, which are not successful either. In the later works of [11, 7, 19], researchers were able to utilize larger corpuses and higher computation capability to expand the original query with all possible related terms, usually in hundreds or thousands, without restricting themselves to terms from certain rigidly grouped equivalence classes. As a result, significant improvements have been reported.

One main weakness of existing automatic expansion methods of [11, 7, 17, 8] is that they depend largely on term co-occurrence to derive terms related to the original query. As a result, these methods tend to append all statistically related terms, or as claimed in their research, synonyms to the original query. In real application, it is often desirable to expand only semantically related terms to create more focused expansion and avoid query topic drifting. Such semantically related terms need to be derived via a domain-specific knowledge source. In this paper, we improve upon the existing automatic query expansion methods by exploring the semantic relationships in a domain-specific knowledge source, and propose a knowledge-based query expansion approach.

In the first phase of our study, we have focused on document retrieval in the medical domain. Recent studies show that real medical queries share great similarity in their structure [4, 5, 6]. Consider the queries in Figure 1 that are taken from a standard test set, OHSUMED [6]. A large number of queries are inquiring about general and semantically related information, e.g., diagnosis and treatment, of a disease. Thus, a query can be represented by a *key concept*, c_{key} (the disease), and several *general supporting concepts*, c_{gs} (e.g. treatment, diagnosis). Usually the general supporting concepts in the query do not match with specific

concepts, e.g, specific treatment methods, in the relevant documents. Consider query #42, "Keratoconus, [1] treatment options." Relevant documents tend to discuss specific methods such as "contact lens" and "penetrating keratoplasty," instead of rephrasing general terms like "treatment" or "option." Simply expanding all the terms that are statistically related will result in an expansion form that contains both treatment-related and non-treatment-related terms. Therefore, to deal with this kind of queries in general, we should restrict the expansion to specific terms that are semantically related, e.g., terms pertinent to the treatment of "Keratoconus." The expanded query can be more focused, understandable to the human user, and yield better retrieval performance.

Query ID	Original Query Form
13	**LACTASE DEFICIENCY** *therapy options*
16	**CHRONIC FATIGUE SYNDROME**, *management and treatment*
37	**FIBROMYALGIA / FIBROSITIS**, *diagnosis and treatment*
38	**DIABETIC GASTROPARESIS,** *treatment.*
42	**KERATOCONUS,** *treatment options.*
43	**BACK PAIN**, *information on diagnosis and treatment*
47	**URINARY RETENTION,** *differential diagnosis*
53	**LUPUS NEPHRITIS**, *diagnosis and management*

Figure 1. Sample OHSUMED queries each containing a key concept and several general supporting concepts. Key concepts are shown in capital letters and general supporting concepts are in italics

To derive semantically related terms from a knowledge source is a challenging task. The semantic relationship between two specific concepts is not usually contained in a knowledge source. This happens because the semantic relationships in all pairs of concepts can be an extremely large set. As a result, it is very unlikely for human experts to identify them one by one. Therefore, we propose to combine term co-occurrence (which can be mined from a corpus) and the Entity-Relation (ER) model (provided by the knowledge source) to automatically derive those semantic relationships.

The rest of this paper is organized as follows. In Section 2, we propose a methodology for knowledge-based query expansion. Section 3 describes the details of deriving semantically related terms, and poses our conjectures that shall be validated experimentally. Section 4 presents the experimental results, and revisits the conjectures with supportive evidence. Section 5 concludes the paper.

2. A FRAMEWORK FOR KNOWLEDGE-BASED QUERY EXPANSION

Figure 2 depicts the knowledge-based query expansion approach. We first derive all statistically correlated terms as candidate expansion terms, whose weights are assigned to reflect their statistical correlation with the original query terms. This step is similar to traditional methods [11]. Second, we use our knowledge-based method to derive a subset of the candidate terms that are semantically related to the original query. This set of semantically related terms, together with their weights assigned in the first step, is then appended to the original query to generate the expanded query.

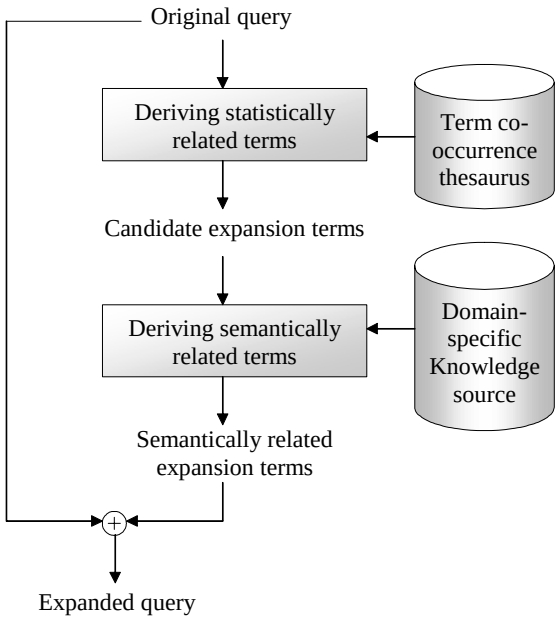

Figure 2. A methodology for knowledge-based query expansion

3. SEMANTICALLY RELATED TERMS IN THE KNOWLEDGE SOURCE

3.1 Method

Because we investigated medical-related corpuses in the first phase of our study, we select UMLS [9], a comprehensive medical knowledge source developed at the National Library of Medicine, as our source of domain-specific knowledge. The core of UMLS consists of three parts: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus contains over 800,000 medical concepts. A group of concepts in the Metathesaurus is abstracted into one semantic type in the Semantic Network. For example, in Figure 3, "Keratoconus" and other disease concepts correspond to one semantic type called "Disease or Syndrome." The Semantic Network is modeled as an Entity-Relation model in which each semantic type is an entity and semantic types are associated via relationships. For example, Figure 3 shows that "Therapeutic and Preventive Procedures," "Medical Device" and "Pharmacologic Substance" are the semantic types that "treats" "Disease or Syndrome." Although UMLS indicates the relationship between semantic types in the Semantic Network level, it does not indicate any relationships among concepts in the Metathesaurus level. For example, UMLS does not provide a "treats" link between a disease concept "Keratoconus" and a medical device concept "Contact Lens."

[1] An eye disease

Therefore, we need a method to automatically derive the semantic relationships among the Metathesaurus concepts, so that for a given key concept c_{key} we can find all the semantically related concepts. The following is the procedure we propose.

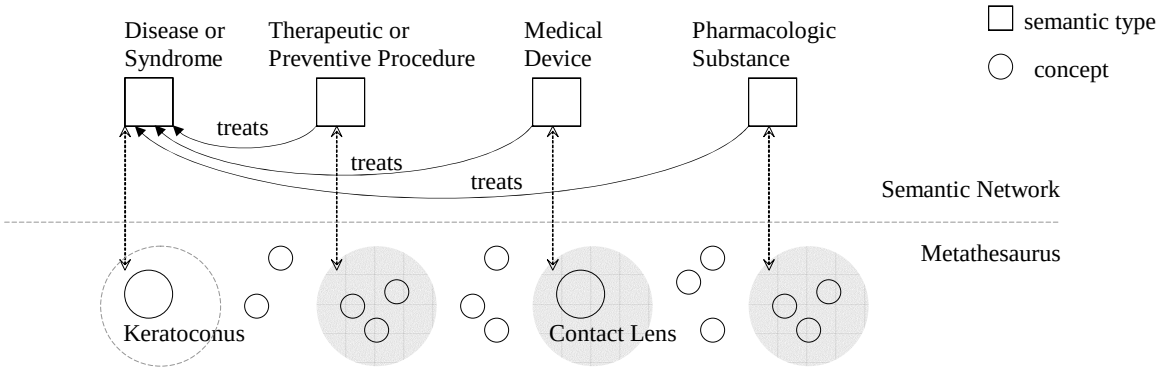

Figure 3. Mapping from concepts to semantic types and the relationships among semantic types

First, we locate the position of c_{key} in the Metathesaurus.

Second, we abstract c_{key} into its corresponding semantic type (e.g. abstracting "Keratoconus" into "Disease or Syndrome").

Third, starting from c_{key}'s semantic type, we follow the relationships as indicated by c_{gs}, e.g., following "treats" if c_{gs} is "treatment options," and reach a set of relevant semantic types.

Fourth, concepts that belong to these relevant semantic types might have more or less the desired relationship with c_{key}. For example, although both belong to "Medical Device," "Contact Lens" is much more likely to be the real treatment option for "Keratoconus" than "CT Scanner." At this step, we do not distinguish among the concepts obtained, and classify them all as semantically related. We shall explain how we make that distinction at the end of this subsection. For the sample query "Keratoconus, treatment options," we reach the shaded circular areas in Figure 3 as the set of semantically related concepts.

Finally, leveraging on the Metathesaurus's concept hierarchy, we add c_{key}'s surrounding concepts to the set of semantically related concepts.

When we have derived the semantically related terms, we intersect them with the candidate expansion terms and use the intersection as the final set of expansion. Recall that each term in the candidate set is assigned a weight according to its statistical co-occurrence with the original query terms. This weight carries over to the final expansion set, to reflect each term's importance in the final expansion. Such weight information distinguishes semantically related terms that highly co-occur with the original query terms, i.e., terms that are truly semantically related, from those that co-occur with the original query terms less often, i.e., terms that belong to the correct semantic type but are marginally relevant to the key concept.

Following the above procedure, we can compute the top (ranked by weight) statistically-related as well as semantically-related word stems, for the query "Keratoconus, treatment options." Figure 4 shows the top 10 stems in each method. It can be seen that the semantically related terms are more pertinent to treatment than the statistically related ones.

Statistically-related expansion stems	Semantically-related expansion stems
Cornea	Cornea Transplantation
Cornea Transplantation	Contact lens
Contact lens	Penetrating keratoplasty
Penetrating keratoplasty	Epikeratoplasty
Epikeratoplasty	Epikeratophakia
Visual Acuities	Eyeglasses
Myopia	Buttons
Epikeratophakia	Radial Keratotomy
Eye	Trephines
Combined corneal dystrophy	Thermokeratoplasty

Figure 4. Top 20 (ranked by weight) statistically-related and semantically-related expansion stems for the query "Keratoconus, treatment options"

4. EXPERIMENTAL RESULTS

4.1 The test collection

Our experiment is based on OHSUMED [6] which is a large medical corpus used in many IR system evaluations. The test set consists of a document collection, a query collection, and a set of relevance judgments.

The document collection is a subset of the MEDLINE database. Each document contains a title, an optional abstract, a set of MeSH headings, author information, publication type, source, a MEDLINE identifier, and a sequence identifier. The query collection consists of 106 queries. Each query contains a patient description, an information request, and a sequence identifier. We use the title, the abstract, and the MeSH heading sub-portions to represent each document, and the information request sub-portion to represent each query.

To testify the effectiveness of the knowledge-based query expansion, we focused on all of the queries in the form of "<key concept>, <general supporting concepts>." There are altogether

49 queries (about half of the total queries) that belong to this set. Further, we excluded 9 queries whose general supporting concepts are not defined by semantic relationships in the UMLS Semantic Network. These excluded general concepts are "complication," "success," "outcome," "prognosis," "admission criteria" and "research" of a particular disease.[2]

For the general supporting concepts in the remaining 40 queries, we asked medical experts in the (removed for blind reviewing) to identify the corresponding semantic relationships in the UMLS Semantic Network. The final identification result is listed in Appendix A.

4.2 Deriving statistically related terms

We select the method in [11] as the AQE method to generate candidate expansion terms.[3] The best automatic query expansion results reported so far are [11], [7] and [19]. The only difference is that during term co-occurrence computation, [11] counts term co-occurrence on an entire document basis, whereas [7] and [19] considers co-occurrence on a local context basis, e.g., when two terms co-occur within a paragraph. In OHSUMED, each document contains at most one abstract which is one paragraph. As a result, the difference between the various methods is insignificant under this experimental setting.

Following the method in [11], we first generate an inverted document vector representation for each term, and then compute pair-wise term co-occurrence using the standard vector dot product. The candidate expansion terms are those having non-zero co-occurrence values with the original query terms. The weight assigned to each expansion term is the average co-occurrence value between this term and each original query term.

4.3 Deriving semantically related terms

After we have derived the set of candidate expansion terms, we follow the general procedure in 3.1 to derive semantically-related terms for each query. We append this set of terms into the original query, together with the weights computed in 4.2, to generate the final expanded query.

4.4 An improved Vector-Space-Model (VSM)

Our knowledge-based method derives concepts that are semantically related to the key concept in the original query. Thus, it is natural to represent both the expanded query and the documents as vectors of concepts. Afterwards, we might apply a concept-based Vector-Space-Model to compute the similarity between the expanded query and the documents.

However, is has been shown that in general concept-indexing cannot outperform word-stem-indexing in terms of retrieval effectiveness. This leads us to adopting an improved VSM, phrase-based VSM [20]. A phrase is the combination of a concept and the word stems that appear in that concept. In the phrase-based VSM, both the query and the documents are represented as vectors of phrases. As a result, the phrase-based VSM is able to match the query against the documents using both concept similarity and word-stem similarity. This method has been shown to be more effective than both the concept-based VSM and the word-stem-based VSM.

In our experiment, we use the knowledge-based expansion method to generate semantically-related concepts. These concepts are then converted into their phrase representation, and expanded into the original query (which is also represented by phrases). We then use the phrase-based VSM to compute the query-document similarity, and from there derive the precision-recall curve for our knowledge-based expansion method.

To compare the performance of our method against existing query expansion solutions, we implemented the method in [11] which expand word stems into the original query. In this replicated work, we use the stem-based VSM to generate the precision-recall curve.

4.5 Document retrieval results

Figure 5 shows the average precision-recall of different methods over the 40 queries. The curve at the bottom (the straight dotted line with no diamonds) is generated by the stem VSM, without any query expansion. The solid curve just above the bottom line is generated by the statistical expansion that expands word stems (i.e. method in [11]). This curve represents the best performance that traditional techniques can achieve, without using any knowledge source. The dashed line immediately above the stem expansion is produced by the phrase VSM without any expansion, and the highest curve of the four is the knowledge-based expansion method that uses the phrase VSM. Compared to the bottom line of stem VSM without expansion, the knowledge-based expansion method has achieved 33% improvement in retrieval effectiveness; compared to the statistical word-stem expansion, the knowledge-based method has achieved 19% improvement. These results strongly indicate that the knowledge-based expansion is more effective in handling queries in the form of <key concept> + <general supporting concepts>.

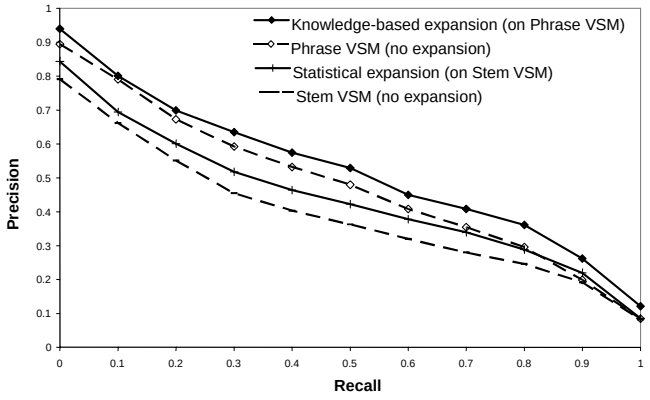

Figure 5. Average precision-recall comparison of the automatic expansion without knowledge and the knowledge-based automatic expansion

5. CONCLUSION

In this paper, we proposed a knowledge-based query expansion to expand semantically related terms into the original query and

[2] The methodology of extending these excluded general supporting concepts into the UMLS Semantic Network will be a future research area

[3] Specifically, [11] use word stems as terms for retrieval

provide more focused query expansion. We used UMLS as the domain-specific knowledge source, and OHSUMED as the test set. Compared to the baseline derived from the original queries without expansion, the knowledge-based expansion yields 9.1% more improvements than the existing statistical correlation-based method. Further, the query expanded by the knowledge-based approach is not only much shorter to save a great amount of computation time, but also more focused and easier for the user to understand. Therefore the knowledge-based query expansion provides significant advantages for document retrieval over the no-knowledge methods, and should be a new direction for future research. We plan to further validate the knowledge-based method with different indexing techniques and with document retrieval in other domain areas.

REFERENCES

[1] C.J. Crouch. An Approach to the Automatic Construction of Global Thesauri. *Information Processing & Management*, 26(5): 629-640, 1990

[2] C.J. Crouch, B. Yong. Experiments in Automatic Statistical Thesaurus Construction. In *Proc. 15th ACM-SIGIR*, pages 77-87, 1992

[3] E.N. Efthimiadis. Query expansion. *Annual Review of Information Science and Technology*, 31:121-187, 1996.

[4] J.W. Ely, J.A. Osheroff, M.H. Ebell, G.R. Bergus, et al. Analysis of questions asked by family doctors regarding patient care. *British Medical Journal*, 319:358-361, 1999.

[5] J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, et al. A taxonomy of generic clinical questions: classification study. *British Medical Journal*, 321:429-432, 2000.

[6] W. Hersh, C. Buckley, T.J. Leone and D. Hickam. OHSUMED: an Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proc. 17th ACM-SIGIR*, pages 191-197, 1994

[7] Y. Jing and W.B. Croft. An association thesaurus for information retrieval. In *Proc. RIAO'94*, pages 146-160, 1994.

[8] R. Mandala, T. Tokunaga, H. Tanaka. Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion, In *Proc. 22nd ACM-SIGIR*, pages 191-197, 1999

[9] National Library of Medicine. *UMLS Knowledge Sources, 12th edition*, 2001.

[10] J.J. Jr. Rocchio. Relevance feedback in information retrieval. In *The Smart system – experiments in automatic document processing*, 313-323, Englewood Cliffs, NJ: Prentice Hall Inc, 1971.

[11] Y. Qiu and H.P. Frei. Concept-based query expansion. In *Proc. 16th ACM-SIGIR*, pages 160-169, 1993.

[12] G. Salton. Relevance feedback and the optimization of retrieval effectiveness. In *The Smart system – experiments in automatic document processing*, 324-336, Englewood Cliffs, NJ: Prentice Hall Inc, 1971.

[13] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*, 1983.

[14] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41(4):288-297, 1990.

[15] K. Sparck Jones. *Automatic keyword classification for information retrieval*. Butterworth, London, 1971.

[16] K. Sparck Jones. Collecting properties influencing automatic term classification. *Information Storage and Retrieval*, 9:499-513, 1973.

[17] E.M. Voorhees. On expanding query vectors with lexically related words. In *Proc. TREC-2*, pages 223-232, 1993.

[18] E.M. Voorhees. Query expansion using lexical-semantic relations. In *Proc. 17th ACM-SIGIR*, pages 61-69, 1994.

[19] J. Xu and W.B. Croft. Query expansion using local and global document analysis. In *Proc. 19th ACM-SIGIR*, pages 4-11, 1996.

[20] W. Mao and W.W. Chu. Free-text medical document retrieval via phrase-based vector space model. In *Proc. AMIA Annual Symposium*, 2002.

APPENDIX: STRUCTURE OF THE 40 STUDIED OHSUMED QUERIES

The following table shows the original forms of the studied queries. To highlight the query structure, the key concepts are capitalized and the general supporting concepts are represented in italics. The general supporting concepts in some of the queries cannot be directly mapped into the relationships in the UMLS Semantic Network. We consult medical experts to translate those general supporting concepts into corresponding UMLS-defined relationships. For example, the general supporting concepts in query #14: "pancytopenia in aids, workup and etiology" are translated into "diagnoses" and "causes" by the medical experts. As a result, the original query becomes the same as "What medical concept **diagnoses** pancytopenia in aids, and what **causes** pancytopenia in aids?"

Query ID	Original Query Form	Relationships in the UMLS Semantic Network that correspond to the query's general supporting concepts
2	*pathophysiology and treatment of* **DISSEMINATED INTRAVASCULAR COAGULATION**	treats, causes
13	**LACTASE DEFICIENCY** *therapy options*	treats
14	**PANCYTOPENIA IN AIDS,** *workup and etiology*	diagnoses, causes
15	**THROMBOCYTOSIS**, *treatment and diagnosis*	treats, diagnoses
16	**CHRONIC FATIGUE SYNDROME**, *management and treatment*	treats
21	**SECONDARY HYPERTENSION**, *recent strategy for workup*	diagnoses

23	**SPONTANEOUS UNILATERAL GALACTORRHEA**, *differential diagnosis and workup*	diagnoses
26	*pathophysiology and etiology of* **STEVENS-JOHNSON SYNDROME**	causes
27	**SICKLE CELL DISEASE**, *treatment advice*	treats
29	**THROMBOCYTOPENIA IN PREGNANCY,** *etiology and management*	treats, causes
32	**COCAINE WITHDRAWAL,** *management*	treats
35	*risk factors and treatment for* **HEPATOCELLULAR CARCINOMA**	treats, causes
Query ID (cont'd)	**Original Query Form**	Relationships in the UMLS Semantic Network that correspond to the query's general supporting concepts
37	**FIBROMYALGIA / FIBROSITIS**, *diagnosis and treatment*	treats, diagnoses
38	**DIABETIC GASTROPARESIS,** *treatment.*	treats
39	**VIRAL GASTROENTERITIS**, *current management*	treats
40	*best treatment* of **MALIGNANT PERICARDIAL EFFUSION IN ESOPHAGEAL CANCER**	treats
41	**ASCITES,** *differential diagnosis and work-up*	diagnoses
42	**KERATOCONUS,** *treatment options.*	treats
43	**BACK PAIN**, *information on diagnosis and treatment*	treats, diagnoses
47	**URINARY RETENTION,** *differential diagnosis*	diagnoses
49	**FLORINEF AND CORONARY ARTERY DISEASE,** *any indications*	indicates
53	**LUPUS NEPHRITIS**, *diagnosis and management*	treats, diagnoses
56	*treatment* of **HYPOTHYROIDISM IN RAPID CYCLING (BIPOLAR DISORDER)**	treats
57	**CEREBRAL EDEMA SECONDARY TO INFECTION**, *diagnosis and treatment*	treats, diagnoses
58	*diagnostic and therapeutic work up* of **BREAST MASS**	treats, diagnoses
64	*prevention, risk factors, pathophysiology of* **HYPOTHERMIA**	causes, prevents
65	**CHRONIC INFLAMMATORY DEMYELINATING POLYNEUROPATHY,** *differential diagnosis and criteria*	diagnoses
67	*outpatient management* of **DIABETES**, *standard management* of **DIABETES** and *any new management techniques*	treats
69	**DIVERTICULITIS**, *differential diagnosis and management*	treats, diagnoses
70	*differential diagnosis* of **ELEVATED ALKALINE PHOSPHATASE AND LDH LEVELS**	diagnoses
72	**THYROTOXICOSIS**, *diagnosis and management*	treats, diagnoses
74	**NEUROLEPTIC MALIGNANT SYNDROME**, *differential diagnosis, treatment*	treats, diagnoses
76	**RADIATION INDUCED THYROIDITIS**, *differential diagnosis, management*	treats, diagnoses
80	**ADRENAL MASS**, *how to work up*	diagnoses
81	**CULTURE NEGATIVE ENDOCARDITIS**, *organisms, diagnosis, treatment*	treats, diagnoses
82	**AIDS DEMENTIA**, *workup*	diagnoses

85	**RECURRENT CELLULITIS,** *risk factors, management, prophylaxis*	treats, causes, prevents
93	**ALLERGIC REACTION TO COUMADIN**, *treatment*	treats
97	**IRON DEFICIENCY ANEMIA**, *which test is best*	diagnoses
98	**SCHEURMANN'S DISEASE**, *treatment*	treats