US 20090281900A1

(19) **United States**

(12) **Patent Application Publication**

(10) **Pub. No.: US 2009/0281900 A1**

(43) **Pub. Date: Nov. 12, 2009**

Rezaei et al.

(54) **DISCOVERING RELEVANT CONCEPT AND CONTEXT FOR CONTENT NODE**

(75) Inventors: **Behnam Attaran Rezaei**, Santa Clara, CA (US); **Riccardo Boscolo**, Culver City, CA (US); **Vwani P. Roychowdhury**, Los Angles, CA (US)

Correspondence Address:
NIXON PEABODY, LLP
401 9TH STREET, NW, SUITE 900
WASHINGTON, DC 20004-2128 (US)

(73) Assignee: **NETSEER, INC.**, Santa Clara, CA (US)

(21) Appl. No.: **12/436,748**

(22) Filed: **May 6, 2009**

Related U.S. Application Data

(60) Provisional application No. 61/050,958, filed on May 6, 2008.

Publication Classification

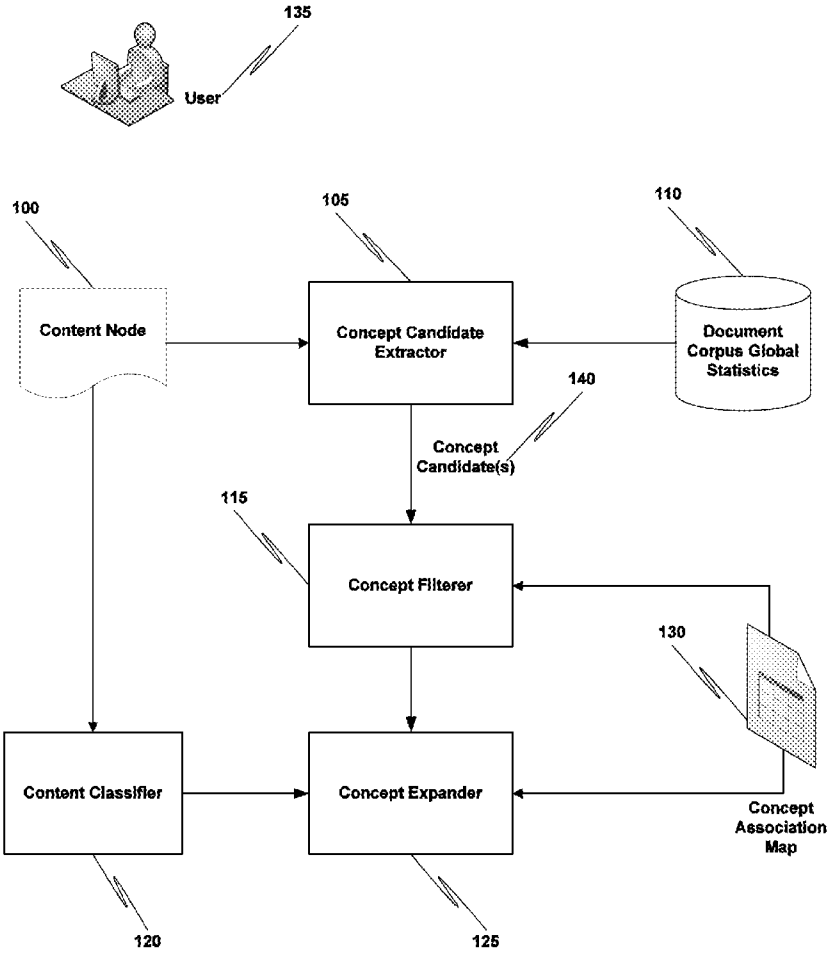
(51) **Int. Cl.**

G06F 17/30	(2006.01)
G06F 15/18	(2006.01)
G06Q 30/00	(2006.01)

(52) **U.S. Cl.** **705/14.55**; 706/12; 707/5; 707/E17.109

(57) **ABSTRACT**

Discovering relevant concepts and context for content nodes to determine a user's intent includes identifying one or more concept candidates in a content node based at least in part on one or more statistical measures, and matching concepts in a concept association map against text in the content node. The concept association map represents concepts, concept metadata, and relationships between the concepts. The one or more concept candidates are ranked to create a ranked one or more concept candidates based at least in part on a measure of relevance. The ranked one or more concept candidates is expanded according to one or more cost functions. The expanded set of concepts is stored in association with the content node.



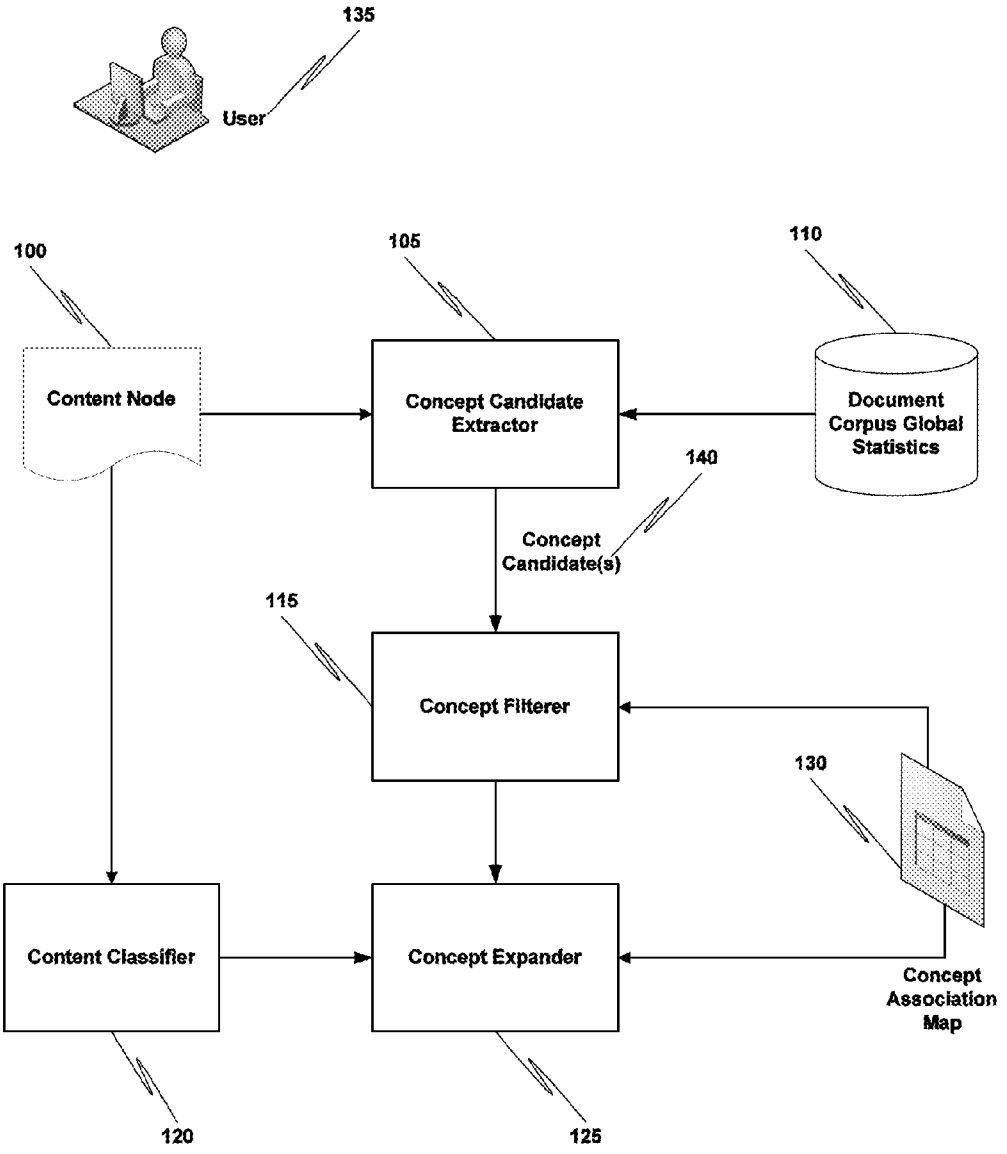
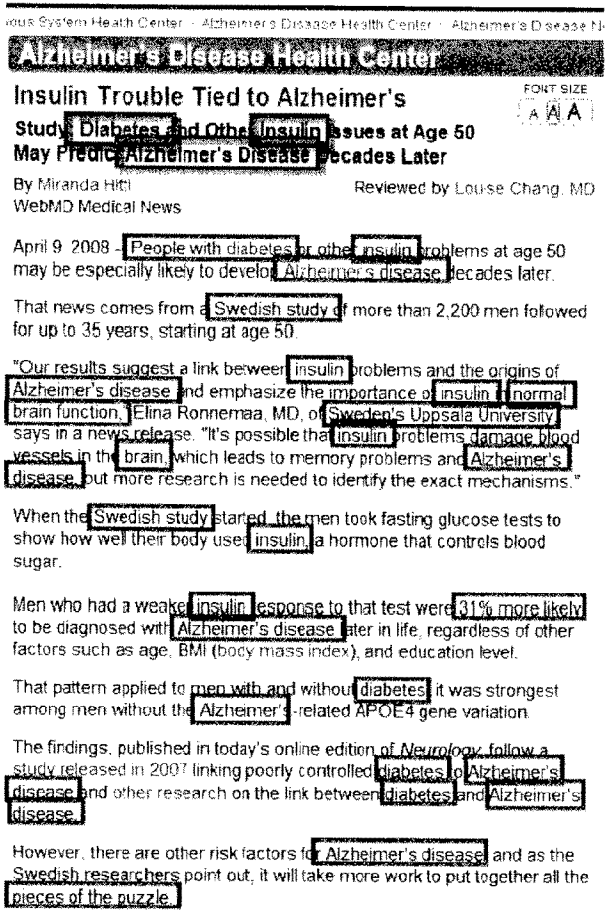

135
User
100
105
110
Content Node
Concept Candidate Extractor
Document Corpus Global Statistics
140
Concept Candidate(s)
115
Concept Filterer
130
Content Classifier
Concept Expander
Concept Association Map
120
125

FIG. 1



Alzheimer's Disease Health Center

Insulin Trouble Tied to Alzheimer's

FONT SIZE A A A

Study: Diabetes and Other Insulin Issues at Age 50 May Predict Alzheimer's Disease Decades Later

By Miranda Hitti
WebMD Medical News

Reviewed by Louise Chang, MD

April 9, 2008 -- People with diabetes or other insulin problems at age 50 may be especially likely to develop Alzheimer's disease decades later.

That news comes from a Swedish study of more than 2,200 men followed for up to 35 years, starting at age 50.

"Our results suggest a link between insulin problems and the origins of Alzheimer's disease and emphasize the importance of insulin in normal brain function," Elina Ronnemaa, MD, of Sweden's Uppsala University, says in a news release. "It's possible that insulin problems damage blood vessels in the brain, which leads to memory problems and Alzheimer's disease, but more research is needed to identify the exact mechanisms."

When the Swedish study started, the men took fasting glucose tests to show how well their body used insulin, a hormone that controls blood sugar.

Men who had a weaker insulin response to that test were 31% more likely to be diagnosed with Alzheimer's disease later in life, regardless of other factors such as age, BMI (body mass index), and education level.

That pattern applied to men with and without diabetes; it was strongest among men without the Alzheimer's-related APOE4 gene variation.

The findings, published in today's online edition of *Neurology*, follow a study released in 2007 linking poorly controlled diabetes to Alzheimer's disease and other research on the link between diabetes and Alzheimer's disease.

However, there are other risk factors for Alzheimer's disease, and as the Swedish researchers point out, it will take more work to put together all the pieces of the puzzle.

Fig 2

Candidate concept extractor

Seed nodes and normalized score:

- Alzheimer's disease, score= 1.0
- diabetes, score= 0.937376
- insulin, score= 0.8725628
- Alzheimer, score= 0.45225
- Swedish study, score= 0.44467
- normal brain function, score= 0.34265
- brain, score= 0.2345652
- study, score= 0.1870422
- people with diabetes, 0.156367
- Sweden's Uppsala University, score= 0.11211
- 31% more likely, score= 0.0389327
- pieces of the puzzle, score= 0.0001111

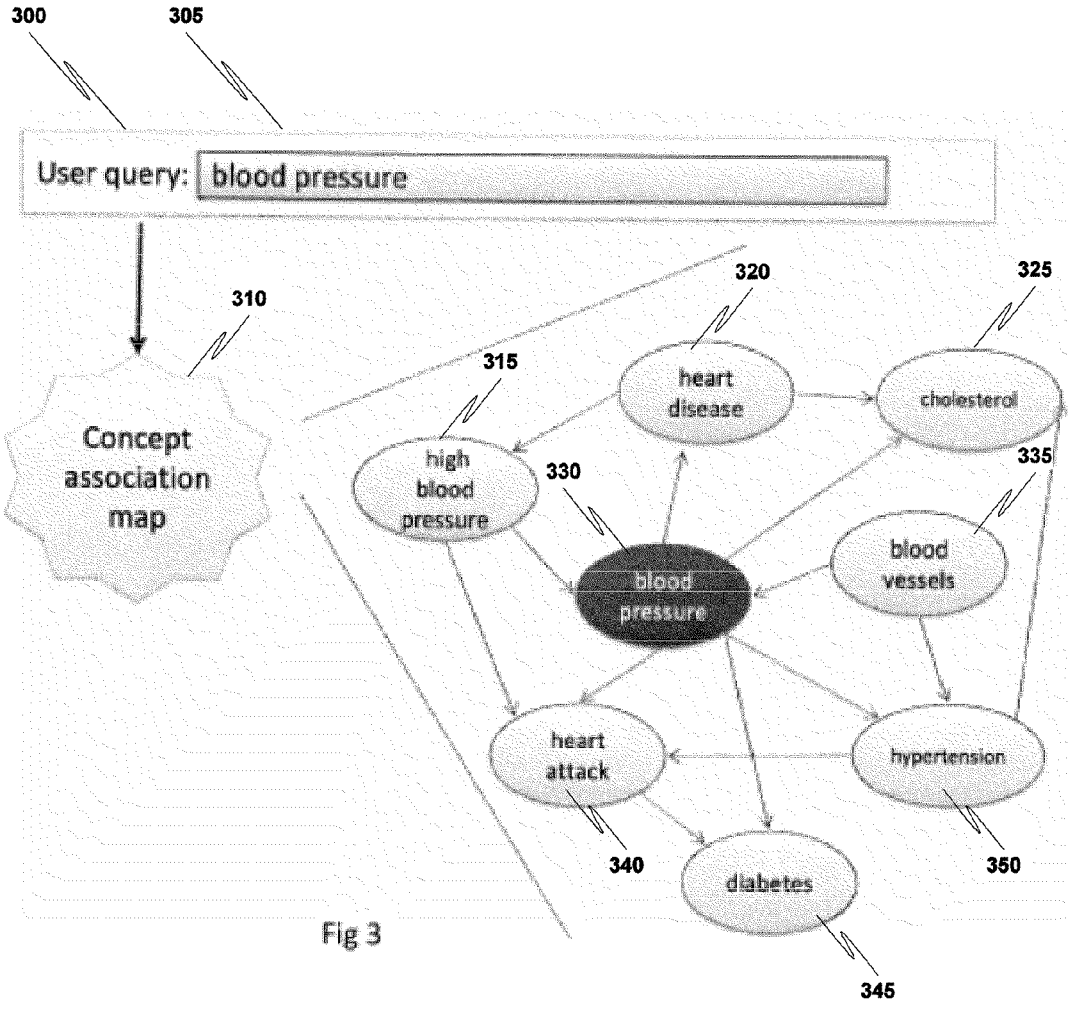

300
305
User query: blood pressure
310
Concept association map
320
325
315
heart disease
cholesterol
high blood pressure
330
335
blood vessels
blood pressure
heart attack
hypertension
340
350
diabetes
345

Fig 3

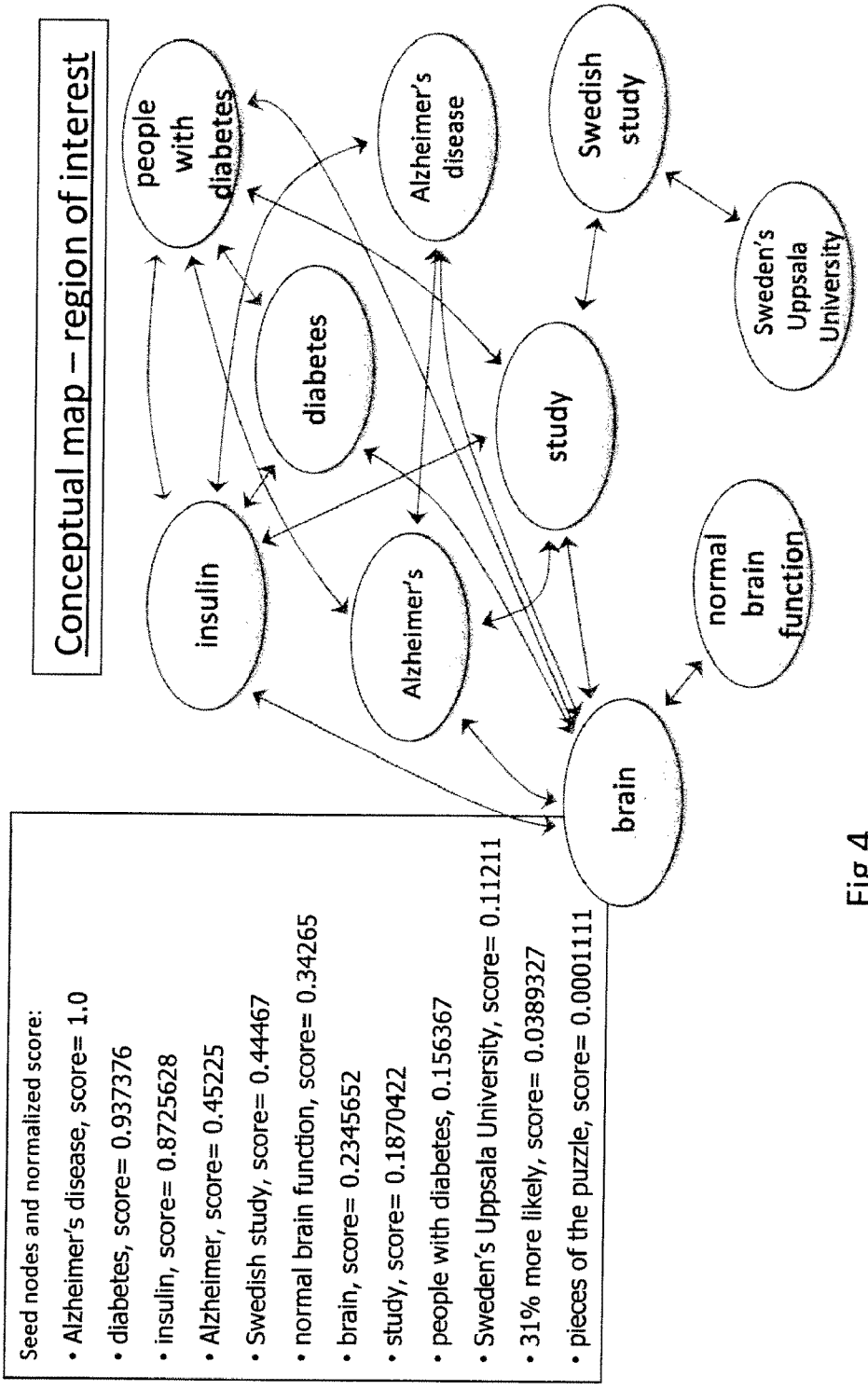

Conceptual map – region of interest
people with diabetes
Alzheimer's disease
Swedish study
insulin
diabetes
study
Sweden's Uppsala University
Alzheimer's
normal brain function
brain
Seed nodes and normalized score:
• Alzheimer's disease, score= 1.0
• diabetes, score= 0.937376
• insulin, score= 0.8725628
• Alzheimer, score= 0.45225
• Swedish study, score= 0.44467
• normal brain function, score= 0.34265
• brain, score= 0.2345652
• study, score= 0.1870422
• people with diabetes, 0.156367
• Sweden's Uppsala University, score= 0.11211
• 31% more likely, score= 0.0389327
• pieces of the puzzle, score= 0.0001111

Fig 4

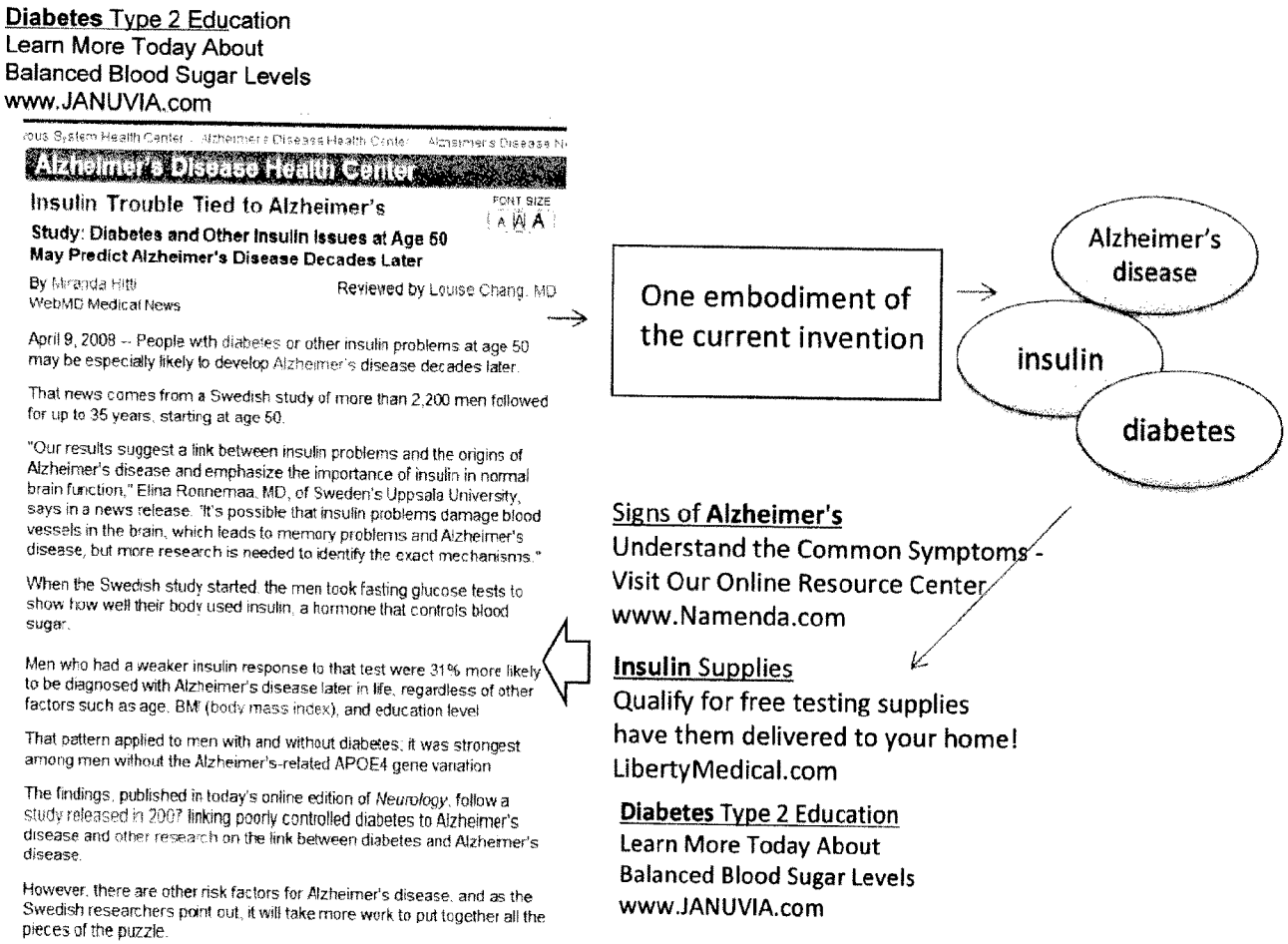

Diabetes Type 2 Education
Learn More Today About
Balanced Blood Sugar Levels
www.JANUVIA.com
Alzheimer's Disease Health Center
Insulin Trouble Tied to Alzheimer's
FONT SIZE
Study: Diabetes and Other Insulin Issues at Age 50
May Predict Alzheimer's Disease Decades Later
By Miranda Hitti
WebMD Medical News
Reviewed by Louise Chang, MD
April 9, 2008 -- People wth diabetes or other insulin problems at age 50 may be especially likely to develop Alzheimer's disease decades later.
That news comes from a Swedish study of more than 2,200 men followed for up to 35 years, starting at age 50.
"Our results suggest a link between insulin problems and the origins of Alzheimer's disease and emphasize the importance of insulin in normal brain function," Elina Ronnemaa, MD, of Sweden's Uppsala University, says in a news release. "It's possible that insulin problems damage blood vessels in the brain, which leads to memory problems and Alzheimer's disease, but more research is needed to identify the exact mechanisms."
When the Swedish study started, the men took fasting glucose tests to show how well their body used insulin, a hormone that controls blood sugar.
Men who had a weaker insulin response to that test were 31% more likely to be diagnosed with Alzheimer's disease later in life, regardless of other factors such as age, BMI (body mass index), and education level.
That pattern applied to men with and without diabetes; it was strongest among men without the Alzheimer's-related APOE4 gene variation.
The findings, published in today's online edition of Neurology, follow a study released in 2007 linking poorly controlled diabetes to Alzheimer's disease and other research on the link between diabetes and Alzheimer's disease.
However, there are other risk factors for Alzheimer's disease, and as the Swedish researchers point out, it will take more work to put together all the pieces of the puzzle.
One embodiment of the current invention
Alzheimer's disease
insulin
diabetes
Signs of Alzheimer's
Understand the Common Symptoms -
Visit Our Online Resource Center
www.Namenda.com
Insulin Supplies
Qualify for free testing supplies
have them delivered to your home!
LibertyMedical.com
Diabetes Type 2 Education
Learn More Today About
Balanced Blood Sugar Levels
www.JANUVIA.com

Fig 5

DISCOVERING RELEVANT CONCEPT AND CONTEXT FOR CONTENT NODE

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of provisional patent application No. 61/050,958 filed May 6, 2008, entitled "Methods and Apparatus for Discovering Relevant Concept and Context for Content Specific Node."

FIELD OF THE INVENTION

[0002] The present invention relates to the field of computer science. More particularly, the present invention relates to discovering relevant concepts and context for content nodes to determine a user's intent, and using this information to provide targeted advertisement and content.

BACKGROUND

[0003] Information retrieval systems are typically designed to retrieve relevant content from a data repository, based on inputs from users. The user input can be in any of the following example forms: (i) a set of keywords, (ii) single or multiple lists of URLs and domains, and (iii) a set of documents (e.g., text files, HTML pages, or other types of markup language content). A goal of such information retrieval systems is to pull the most relevant content (i.e., most relevant to the given input) from the underlying repository, which might itself consist of a heterogeneous set of structured and unstructured content. An example of the aforementioned information retrieval system is a traditional search engine, where a user provides a set of keywords, and the search engine provides simple ranked lists of top relevant web pages, and a separate list of top relevant paid listings or sponsored links. The set of web pages matching user's search queries and the advertisement database containing sponsored advertising materials are currently two separate databases that are processed very differently to pull the relevant pages and the sponsored links for the same user query. Thus, the conventional search engine described above provides an example of two distinct information repositories being processed in response to the same query.

[0004] Current systems find important keywords of a web page then try to expand them using various resources. This expanded set of keywords is compared with a user-provided set of keywords. One problem with such an approach is that keywords can have different meanings. For example, "Chihuahua" is a dog breed, but it is also a province in Mexico. In current systems, Chihuahua may expand to:

[0005] Chihuahua Breeders,

[0006] Travel to Chihuahua

[0007] Travel to Mexico

[0008] Chihuahua Puppy

[0009] Dog Training

[0010] Hotels in Chihuahua

[0011] Teacup Chihuahua Puppies

[0012] Cheap flights,

[0013] A person interested in a Chihuahua dog would find information about the Chihuahua province or travel to it less useful. And a person interested in the Chihuahua province would find information about dog training or a Chihuahua dog less useful. Without knowing the context of the user-provided set of keywords, current systems often present search results that are irrelevant to what the user is seeking.

[0014] While the aforementioned systems allow for limited targeting of advertisement and content, such systems fail to provide efficient targeted advertisement avenues. Accordingly, a need exists for an improved solution for advertisement targeting.

SUMMARY

[0015] The content in a content node is expanded into groupings of concepts and phrases, where each such group represents one possible user intention (as implied by the query phrase or keyword). Each such grouping is analyzed to provide relevant content, such as unstructured data like World Wide Web data, categorized data, display advertisements, and paid listings. This more accurately reflects user intentions even for cases where click through information is absent.

[0016] A computerized system for finding important keywords on a content node uses its content and other related URLs like domains. The system is capable of clustering and pruning them by projecting such keywords and phrases on a predefined conceptual map. The projection on the conceptual map enables the expansion of the user intention into multiple contexts, and the further identification of content relevant to the original content node.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The accompanying drawings, which are incorporated into and constitute a part of this specification, illustrate one or more embodiments of the present invention and, together with the detailed description, serve to explain the principles and implementations of the invention.

[0018] In the drawings:

[0019] FIG. **1** is a block diagram that illustrates of a system for discovering relevant concepts and context for content nodes in accordance with one embodiment of the present invention.

[0020] FIG. **2** is a block diagram that illustrates extracting candidate seed concepts from a document in accordance with one embodiment of the present invention.

[0021] FIG. **3** is a block diagram that illustrates conceptual expansion of seed nodes onto a concept map in accordance with one embodiment of the present invention.

[0022] FIG. **4** is a block diagram that illustrates an example of a region of interest in concept space for a particular input page in accordance with one embodiment of the present invention.

[0023] FIG. **5** is a block diagram that illustrates matching pay per click (PPC) advertisements to web pages in accordance with one embodiment of the present invention.

DETAILED DESCRIPTION

[0024] Embodiments of the present invention are described herein in the context of discovering relevant concepts and context for content nodes to determine a user's intent, and using this information to provide targeted advertisement and content. Those of ordinary skill in the art will realize that the following detailed description of the present invention is illustrative only and is not intended to be in any way limiting. Other embodiments of the present invention will readily suggest themselves to such skilled persons having the benefit of this disclosure. Reference will now be made in detail to implementations of the present invention as illustrated in the accompanying drawings. The same reference indicators will

be used throughout the drawings and the following detailed description to refer to the same or like parts.

[0025] The invention examines content of interest to a user, to determine what concepts are most closely associated with that content. Other content that is closely associated with the same concepts taken in context is more likely be of interest to the user. And other content that has similar words but different concepts is less likely be of interest to the user. The invention uses concept information previously gleaned from an analysis of other web pages to better understand the context of a current web page. Concepts extracted from the current web page that are not related to the current context are pruned. The content known to be of interest to the user may be presented along with other content that is closely associated with the concepts related to the current context, thus increasing the likelihood that the user will find the other content interesting.

[0026] For example, suppose a user visits a web page describing the "Chihuahua" province of Mexico. The "Chihuahua" may expand to:

[0027] Chihuahua Breeders,

[0028] Travel to Chihuahua

[0029] Travel to Mexico

[0030] Chihuahua Puppy

[0031] Dog Training

[0032] Hotels in Chihuahua

[0033] Teacup Chihuahua Puppies

[0034] Cheap flights

[0035] But the current context relates to the "Chihuahua" province, not the Chihuahua dog breed. According to the invention, concepts extracted from the current web page that are not related to the current context are pruned, resulting in only concepts related to the current context:

[0036] Travel to Chihuahua

[0037] Travel to Mexico

[0038] Hotels in Chihuahua

[0039] Cheap flights

The current web page may be presented along with other content (e.g. paid listings or other websites) that is closely associated with these four concepts that are related to the current context, thus increasing the likelihood that the user will find the other content interesting.

[0040] In the context of the present invention, the term "content node" refers to one or more groupings of data. Example groupings of data include a web page, a paid listing, a search query, and a text file.

[0041] In the context of the present invention, the term "concept" refers to a unit of thought, expressed by a term, letter, or symbol. It may be the mental representation of beings or things, qualities, actions, locations, situations, or relations. A concept may also arise from a combination of other concepts. Example concepts include "diabetes," "heart disease," "socialism," and "global warming."

[0042] In the context of the present invention, the term "concept association map" refers to a representation of concepts, concept metadata, and relationships between the concepts.

[0043] FIG. **1** is a block diagram that illustrates a system for discovering relevant concepts and context for content nodes in accordance with one embodiment of the present invention. As shown in FIG. **1**, concept association map **130** includes concepts and their relationships, which may be expressed as bi-directional edges. Concepts are nodes in a graph, and different kinds of meta-data are associated with each such node. For example, the node meta-data can include the frequency of appearance of the concept in a given corpus, its structural relevance in the graph, cost per action (CPA) and click through rate (CTR) data for ads associated with it, CTR data for the concept itself as derived from user **135** browsing patterns, as well as a labeling that associates it with a specific category. Unlike static concept databases, this concept association map **130** is dynamic and it is continuously updated by the system.

[0044] According to one embodiment of the present invention, the concept association map **130** is derived from different sources. Example sources include concept relationships found on the World Wide Web, associations derived from users **135** browsing history, advertisers bidding campaigns, taxonomies, and encyclopedias.

[0045] Still referring to FIG. **1**, concept candidate extractor **105** is configured to identify one or more relevant concept candidates in a content node **100**. Concept candidate extractor **105** relies at least in part on a set of statistical measures (document corpus global statistics **110**) in order to identify such candidates. According to one embodiment of the present invention, one or more of the following statistical measures **110** are used to extract concept candidates:

[0046] a. Global document frequency of n-grams defining a concept. This measure is indicative of the likelihood that a given n-gram will appear in a document that is part of a corpus.

[0047] b. Frequency of n-grams in the content node **100**.

[0048] c. Similarity of the content node **100** to other content nodes for which relevant concept candidates have already been identified.

[0049] d. Weight of the node in the concept graph.

[0050] According to one embodiment of the present invention, concept candidates **140** are extracted from different input sources associated with a page on the World Wide Web, viz. the body of the HTML page, the title, the meta-data tags, the anchor text of hyperlinks pointing to this page, the anchor text of hyperlinks contained in the page, the publishing history of the page, as well as the same type of input sources for pages related to this one.

[0051] According to one embodiment of the present invention, the content to be tagged with concepts is provided directly by the user **135**, for example in the form of a text file.

[0052] According to one embodiment of the present invention, the content to be tagged is any textual section of a relational database, e.g. a product inventory database.

[0053] According to another embodiment of the present invention, the node content is a user query, defined as a set of search keywords.

[0054] According to another embodiment of the present invention, the concept candidates **140** are provided by the user **135** as input to the system. For example, in a bidding campaign a content provider or a merchant could provide such a list based on internal knowledge about the products to be advertised.

[0055] According to one embodiment of the present invention, for web page, top referral queries on major search engines are also identified as top concepts. For example, if for a URL a.b.com/d, most of the incoming traffic from major search engines is coming from users **135** searching for query "diabetes" and "diabetes symptoms," these queries are added as top concepts.

[0056] According to another embodiment of the present invention, concepts can also get identified from other pages relevant to the page of interest, for example if the relevant

page is structurally similar (through hyperlinks) to the page of interest, or if the relevant page is contextually similar (same content) to the page of interest.

[0057] Concept candidate extractor **105** is configured to use the aforementioned statistics to extract suitable concept candidates in the content node **100**. This is accomplished by matching the concepts available in the concept association map **130** against the text in the content node **100**.

[0058] Concept filterer **115** is configured to rank the concept candidates **140** based at least in part on a measure of relevance that weighs their frequency in the content node **100**, their likelihood of appearing in a document, as well as the likelihood of being selected based on the closeness of this content node **100** to similar concept nodes.

[0059] According to another embodiment of the present invention, for the case of structured content (e.g. a web page), different content sections are weighed according to their relative importance. For example, the title of a page is weighted more than the body of the page, relative to its length.

[0060] FIG. **2** is a block diagram that illustrates extracting candidate seed concepts from a document in accordance with one embodiment of the present invention. As shown in FIG. **2**, node content is: [type=web page, url=http://www.webmd.com/alzheimers/news/20080409/insulin-trouble-tied-to-alzheimers]

[0061] The concept candidates selected and their respective scores are, ranked in order of decreasing relevance:

[0062] [insulin, score=1.0]

[0063] [diabetes, score=1.0]

[0064] [people with diabetes, score=0.873529]

[0065] [Swedish study, score=0.123344]

[0066] [Alzheimer's disease, score=0.3456]

[0067] [study, score=0.43222]

[0068] [brain, score=0.986292]

[0069] [normal brain function, score=0.563738]

[0070] [more research is needed, score=0.23445]

[0071] [Sweden's Uppsala University, score=0.432122]

[0072] [released in 2007, score=0.13456]

[0073] [31% more likely to be diagnosed, score=0.11111]

[0074] [pieces of the puzzle, score=0.0045466]

[0075] The mapping of the candidates against the available conceptual map shows that the following concept candidates are associated with high score relative to other concept candidates: [diabetes, people with diabetes, Alzheimer's disease, brain, normal brain function, insulin].

[0076] Referring again to FIG. **1**, concept expander **125** is configured to expand the initial set of seed concepts by selecting neighbors of such seed nodes, according to a set of cost functions. This is described in more detail below, with reference to FIG. **3**.

[0077] FIG. **3** is a block diagram that illustrates conceptual expansion of seed nodes onto a concept map in accordance with one embodiment of the present invention. In FIG. **3**, the user has entered "blood pressure." "Blood pressure" is found in concept association map **310** and is thus a seed concept. As shown in the blown-up portion of concept association map **310**, the neighbors of the "blood pressure" concept **330** are "high blood pressure" **315**, "heart disease" **320**, "cholesterol" **325**, "blood vessels" **335**, "hypertension" **350**, "diabetes" **345**, and "heart attack" **340**.

[0078] One cost function is based on selecting the neighbors that present the best clustering characteristics, i.e. they are more likely to be strongly associated with each other.

[0079] Another cost function is based on selecting neighbors that, based on aggregate user activity history, have a higher likelihood to be associated.

[0080] Another cost function is based on looking at the likelihood that such concepts are selected together based on their co-occurrence in a corpus of documents (e.g. the World Wide Web).

[0081] Another cost function is based on determining which neighboring concepts in the concept association map are tied to a form of monetization (e.g. online advertisement) that yields the highest conversion rate (measured as a combination of CPC and CTR).

[0082] According to one embodiment of the present invention, the nodes in the concept association map are also tagged with labels representing one or more high-level categories.

[0083] According to one embodiment of the present invention, a page content classifier **120** is utilized to label the page with a high level category in order to narrow down the mapping to the concept association map **130** to certain pre-defined contexts.

[0084] According to one embodiment of the present invention, results on the concept association map **130** are clustered to identify different user's intention.

[0085] According to another embodiment of the present invention, the highest-weighted concepts in the graph are chosen as top related concepts. Weight score can be defined using different sources. Examples of weights to be used are structural scores like "betweenness" and "page rank," monetization values like click through and cost per click and frequency of appearance on the web or user's query logs.

[0086] FIG. **4** is a block diagram that illustrates an example of a region of interest in concept space for a particular input page in accordance with one embodiment of the present invention. The concept association map is augmented by adding links between search queries and concepts on the concept association map through the pages that received these referrals. For example, if page a.b.c.com/d receives a large volume of search traffic for the term "diabetes diagnostic" and the page is mapped to region of interest with top concepts: "diabetes disease," "diabetes symptoms," and "type 2 diabetes," the weight of the link between these concepts and "diabetes diagnostic" is increased.

[0087] According to another embodiment of the present invention, top concepts and regions of interest are used to map paid listings or other forms of advertisement to the content node as described in FIG. **5**. FIG. **5** is a block diagram that illustrates matching pay per click (PPC) advertisements to web pages in accordance with one embodiment of the present invention. In FIG. **5**, a web page displaying a document entitled "Insulin Trouble Tied to Alzheimer's" is shown on the left, and three paid listings are shown on the right. The document and the paid listings are considered content nodes. The top concepts identified in the document are "Alzheimer's disease," "insulin," and "diabetes." The particular three paid listings are selected based on how closely the listings are identified with the same top concepts identified in the document. The paid listing "Signs of Alzheimer's" is identified with the concept "Alzheimer's disease." The paid listing "Insulin Supplies" is identified with the concept "insulin." The paid listing "Diabetes Type 2 Education" is identified with the concept "diabetes." The effectiveness of the paid listings is increased by placing the listings near the document identified with the same concepts.

[0088] In the interest of clarity, not all of the routine features of the implementations described herein are shown and described. It will, of course, be appreciated that in the development of any such actual implementation, numerous implementation-specific decisions must be made in order to achieve the developer's specific goals, such as compliance with application- and business-related constraints, and that these specific goals will vary from one implementation to another and from one developer to another. Moreover, it will be appreciated that such a development effort might be complex and time-consuming, but would nevertheless be a routine undertaking of engineering for those of ordinary skill in the art having the benefit of this disclosure.

[0089] According to one embodiment of the present invention, the components, process steps, and/or data structures may be implemented using various types of operating systems (OS), computing platforms, firmware, computer programs, computer languages, and/or general-purpose machines. The method can be run as a programmed process running on processing circuitry. The processing circuitry can take the form of numerous combinations of processors and operating systems, connections and networks, data stores, or a stand-alone device. The process can be implemented as instructions executed by such hardware, hardware alone, or any combination thereof. The software may be stored on a program storage device readable by a machine.

[0090] While embodiments and applications of this invention have been shown and described, it would be apparent to those skilled in the art having the benefit of this disclosure that many more modifications than mentioned above are possible without departing from the inventive concepts herein. The invention, therefore, is not to be restricted except in the spirit of the appended claims.

What is claimed is:

1. A computerized method comprising:

identifying one or more concept candidates in a content node based at least in part on:

one or more statistical measures; and

matching concepts in a concept association map against text in the content node, the concept association map representing concepts, concept metadata, and relationships between the concepts;

ranking the one or more concept candidates to create a ranked one or more concept candidates based at least in part on a measure of relevance;

expanding the ranked one or more concept candidates according to one or more cost functions, the expanding creating an expanded set of concepts; and

storing the expanded set of concepts in association with the content node.

2. The method of claim **1**, wherein the one or more statistical measures comprises one or more of:

an indication of a likelihood that a given n-gram will appear in a document that is part of a corpus;

a frequency of n-grams in the content node;

a similarity of the content node to other content nodes for which relevant concept candidates have already been identified; and

a weight of the content node in the concept association map.

3. The method of claim **1**, wherein the measure of relevance weighs:

a frequency that the one or more concept candidates occurs in the content node;

a likelihood of the one or more concept candidates appearing in a document; and

a likelihood of the one or more concept candidates being selected based on the closeness of the content node to similar concept nodes.

4. The method of claim **1**, further comprising:

before the identifying, classifying the content of the content node.

5. The method of claim **1**, wherein the concept association map is derived from one or more of:

concept relationships found on the World Wide Web;

associations derived from user browsing history;

advertisers bidding campaigns;

taxonomies; and

encyclopedias.

6. The method of claim **1**, wherein the content node comprises a user query having a set of search keywords.

7. The method of claim **1**, wherein the one or more concept candidates are provided by a user.

8. The method of claim **1**, wherein the one or more cost functions comprises selecting neighbors that are more likely to be strongly associated with each other.

9. The method of claim **1**, wherein the one or more cost functions comprises selecting neighbors that, based on aggregate user activity history, have a higher likelihood to be associated.

10. The method of claim **1**, wherein the one or more cost functions comprises determining which neighboring concepts in the concept association map are tied to a form of monetization that yields the highest conversion rate.

11. The method of claim **1**, further comprising mapping one or more advertisements to the content node based at least in part on the expanded set of concepts.

12. An apparatus comprising:

a concept association map representing concepts, concept metadata, and relationships between the concepts;

a candidate concept extractor configured to identify one or more concept candidates in a content node based at least in part on:

one or more statistical measures; and

matching concepts in a concept association map against text in the content node;

a concept filterer configured to rank the one or more concept candidates to create a ranked one or more concept candidates based at least in part on a measure of relevance; and

a concept expander configured to expand the ranked one or more concept candidates according to one or more cost functions, the expanding creating an expanded set of concepts, the apparatus further configured to store the expanded set of concepts in association with the content node.

13. The apparatus of claim **12**, wherein the one or more statistical measures comprises one or more of:

an indication of a likelihood that a given n-gram will appear in a document that is part of a corpus;

a frequency of n-grams in the content node;

a similarity of the content node to other content nodes for which relevant concept candidates have already been identified; and

a weight of the content node in the concept association map.

14. The apparatus of claim **12**, wherein the measure of relevance weighs:

a frequency that the one or more concept candidates occurs in the content node;

a likelihood of the one or more concept candidates appearing in a document; and

a likelihood of the one or more concept candidates being selected based on the closeness of the content node to similar concept nodes.

15. The apparatus of claim **12**, wherein the apparatus is further configured to, before the identifying, classify the content of the content node.

16. The apparatus of claim **12**, wherein the concept association map is derived from one or more of:

concept relationships found on the World Wide Web;

associations derived from user browsing history;

advertisers bidding campaigns;

taxonomies; and

encyclopedias.

17. The apparatus of claim **12**, wherein the content node comprises a user query having a set of search keywords.

18. The apparatus of claim **12**, wherein the one or more concept candidates are provided by a user.

19. The apparatus of claim **12**, wherein the concept expander is further configured to select neighbors that are more likely to be strongly associated with each other.

20. The apparatus of claim **12**, wherein the concept expander is further configured to select neighbors that, based on aggregate user activity history, have a higher likelihood to be associated.

21. The apparatus of claim **12**, wherein the concept expander is further configured to determine which neighboring concepts in the concept association map are tied to a form of monetization that yields the highest conversion rate.

22. The apparatus of claim **12**, wherein the apparatus is further configured to map one or more advertisements to the content node based at least in part on the expanded set of concepts.

23. An apparatus comprising:

means for identifying one or more concept candidates in a content node based at least in part on:

one or more statistical measures; and

matching concepts in a concept association map against text in the content node, the concept association map representing concepts, concept metadata, and relationships between the concepts;

means for ranking the one or more concept candidates to create a ranked one or more concept candidates based at least in part on a measure of relevance;

means for expanding the ranked one or more concept candidates according to one or more cost functions, the expanding creating an expanded set of concepts; and

Means for storing the expanded set of concepts in association with the content node.

24. A program storage device readable by a machine, embodying a program of instructions executable by the machine to perform a method, the method comprising:

identifying one or more concept candidates in a content node based at least in part on: one or more statistical measures; and

matching concepts in a concept association map against text in the content node, the concept association map representing concepts, concept metadata, and relationships between the concepts;

ranking the one or more concept candidates to create a ranked one or more concept candidates based at least in part on a measure of relevance;

expanding the ranked one or more concept candidates according to one or more cost functions, the expanding creating an expanded set of concepts; and

storing the expanded set of concepts in association with the content node.

* * * * *