Scientometrics (2015) 102:905–927
DOI 10.1007/s11192-014-1372-x

Identifying the landscape of Alzheimer's disease research with network and content analysis

Min Song · Go Eun Heo · Dahee Lee

Received: 31 March 2014 / Published online: 17 July 2014

Abstract Alzheimer's disease (AD) is one of degenerative brain diseases, whose cause is hard to be diagnosed accurately. As the number of AD patients has increased, researchers have strived to understand the disease and develop its treatment, such as medical experiments and literature analysis. In the area of literature analysis, several traditional studies analyzed the literature at the macro level like author, journal, and institution. However, analysis of the literature both at the macro level and micro level will allow for better recognizing the AD research field. Therefore, in this study we adopt a more comprehensive approach to analyze the AD literature, which consists of productivity analysis (year, journal/proceeding, author, and Medical Subject Heading terms), network analysis (co-occurrence frequency, centrality, and community) and content analysis. To this end, we collect metadata of 96,081 articles retrieved from PubMed. We specifically perform the concept graph-based network analysis applying the five centrality measures after mapping the semantic relationship between the UMLS concepts from the AD literature. We also analyze the time-series topical trend using the Dirichlet multinomial regression topic modeling technique. The results indicate that the year 2013 is the most productive year and Journal of Alzheimer's Disease the most productive journal. In discovery of the core biological entities and their relationships resided in the AD related PubMed literature, the relationship with glycogen storage disease is founded most frequently mentioned. In addition, we analyze 16 main topics of the AD literature and find a noticeable increasing trend in the topic of transgenic mouse.

Keywords Alzheimer's disease (AD) · Bibliometrics · Document representation · Concept graph · Topic modeling

M. Song (✉) · G. E. Heo · D. Lee
Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Korea
e-mail: min.song@yonsei.ac.kr

Introduction

Alzheimer's disease (AD) is a fatal, progressive brain disease that causes people, particularly elders of age 60 or above, the fatal brain disorders. According to the fact sheets of AD, it is the sixth leading cause of death in the United States (http://www.alz.org/alzheimers_disease_facts_and_figures.asp). Experts estimate that approximately 5.1 million Americans may have AD. In parallel with an increased interest in AD, the research community pays more attention to research on AD. As of April 23rd in 2014, 96,279 articles were retrieved by the query ["Alzheimer disease" OR "Alzheimer's disease"] in PubMed. In PubMed Central, about 67,000 full-text articles were retrieved by the same query.

There are several attempts to map out research on AD with bibliometrics (Chen et al. 2014; Sorensen 2009; Sorensen et al. 2010). Bibliometrics is a type of a research method to utilize quantitative and statistical analyses to describe patterns of publication within a given field or body of literature (Song et al. 2014). Sorensen et al. (2010) conducted author co-citation analysis with 269 Alzheimer investigators and 167,142 researchers to identify major researchers in AD. Bibliometrics analysis of cholinesterase inhibitors was conducted to find the current trend of AD on 4,982 from Science Citation Index Expanded (Chen et al. 2014). Sorensen identified top 100 AD researchers from PubMed and Social Science Index to assess productivity and impact by an AD-specific H-index (Sorensen 2009). Previous studies on understanding the field primarily focus on macro analysis including who are the leading institutes and researchers. However, a more comprehensive investigation of this field is demanded to understand the field both at a macro and a micro level.

The present study aims to identify characteristics of the literature published on AD over five decades from pre-1950 to 2014 (primarily 2000–2013) by examining the topic trends and the network of biological entities on AD. Unlike previous studies, on top of bibliometric analysis including leading researchers in AD, the present paper provides a deeper analysis by examining topical changes over a certain time period and salient biological entities and interactions among them. To this end, first, we propose a Concept Graph approach to identify frequently mentioned entities and the relationships among them with each other. Second, we apply a Topic Modeling technique to understand a topical trend of AD over time.

A graph, representing an article, is constructed from a small set of key concepts that represent the text content or the topics of the documents. The graph is generated using minimal information that is either extracted from the text or made available from other sources such as authors. It can be expanded into higher degrees through mapping external domain knowledge. The features that we consider in this study are biological entities such as gene, protein, and disease expressed in the literature as well as relationships that might exist among them. The extracted entities represent specific and general concepts in medicine, biology, and related fields such as: diseases, anatomical structures, pharmacologic substances, biologic functions, and others. The relationships are also predefined and are of semantic nature and include synonyms, parent–child relationships, sibling relationships, and other narrow or broad relationships defined in the ontology. In particular, we map the initial key-concept list, representing a document, to concepts defined in the unified medical language system (UMLS)[1] (Lindberg et al. 1993)—a comprehensive biomedical ontology and a thesaurus that contains definitions and semantic information of a wide range of

[1] "Unified Medical Language System (UMLS)—Home" [Online]. http://www.nlm.nih.gov/research/umls/. Accessed: 8 Feb 2013.

concepts in medicine, biology, health sciences, and related domains. The resulting document representation is therefore a graph of concept nodes where the edges connecting them represent semantic relationships that exist among the concepts.

We identify topical trends in AD over time. To this end, we adopt Dirichlet multinomial regression (DMR), a variation of latent Dirichlet allocation. DMR topic models condition on observed features of the document, such as author, publication venue, references, and dates (Mimno and McCallum 2008). DMR is a generative topic model to generate both the words and the metadata simultaneously given hidden topic variables. In this paper, DMR topic models condition on year to track topical trends over time.

The rest of the paper is organized in the following order: "Related work" section reviews related works on document representation and bibliometrics; "Proposed approach" section presents the research methods used in this study; "Results and discussion" section analyzes experimental results; "Conclusion" section summarizes the results and provides implications for future research.

Related work

Document representation

A prevalent document representation model is the bag-of-words model where each document is transformed into a collection of terms or words, without taking the order in which they appear in the text or the existence of semantic or other relationships between the words into consideration (Salton et al. 1975). Other similar approaches extend this representation and use n-grams features to represent combinations of characters (Damashek 1995) or words (Cavnar and Trenkle 1994) of a text's content and apply it in classification techniques. The vector space model weighing scheme was also used to represent sentences in a document, as described in (Gong and Liu 2001), where documents are decomposed into sentences and each sentence was represented as a weighted vector of term frequencies and applied in a text summarization application.

Other efforts have also been made to utilize the structure and semantics of the text and incorporate them into the representation to enhance the used techniques. For example, Shehata et al. (2007) incorporated the semantic structure at both sentence and document levels. Their models combined statistical features and a conceptual ontological graph representation that represents the sentence structure while maintaining the sentence semantics in the original document. In the research of Andreasen et al. (2009), the authors transformed documents into a space of conceptual feature structures using ontology and lexical resources for a higher level representation and applied it in content-based search. In the study of Ercan and Cicekli (2007) a lexical chain that holds a set of semantically related words of a document was used to represent the semantic content of a portion of the document. Huang et al. (2006) presented a keywords extraction algorithm. Each document is treated as a semantic network that holds both syntactic and statistical information. A semantic network model was developed in which each term is represented by a node and a relation between two terms by an edge. Additional in-depth description of the use of the vector space model and semantics in capturing meaning of the text as well as their applications can be found in the study of Turney and Pantel (2010).

Graph structures have also been used to represent documents as they preserve the structure embedded in the content and allow using graph techniques that have a strong algorithmic and mathematical foundation in discrete math and computer science. For instance, Chen et al. (2005) proposed a graph representation for document summarization tasks. They use a thesaurus and association rules to connect key phrases in the text. Wan et al. (2007) claimed that graphs are also used to represent documents for summarization. The graphs capture word–word, word-sentence, and sentence–sentence relationships in the text. The word and sentence saliency scores are then computed to rank the results. Similarly, ontology-based mapping of text into concept graphs have been used in text categorization (Bleik et al. 2009) applied on biomedical datasets where the graph features are incorporated into the representation.

Term or keyphrase statistics, such as occurrence frequencies extracted from the text, are usually essential for learning and classification and have been successfully used in text categorization and other text mining applications. However, in this paper we address the problem when such information is not available, perhaps due to the absence or limited availability of the full-text content, or when the documents are very large and using an alternative reduced representation would be desired. The method also highlights how domain knowledge can be incorporated into the representation and applied in text categorization. In the following section we describe the method of representing a text document starting with a few available key concepts that characterize the document.

Bibliometrics in Alzheimer's disease research

There have been a few bibliometric studies in the field of AD. Ansari et al. (2006) applied the basic bibliometric analysis with the different parameters such as the distribution of country, authorship, journal, subject, language, etc. Sorensen (2009) firstly tried a literature analysis through citation analyses and productivity filters. He explained the role of AD within the neruroscience field and the brief summary of many research foci within the AD scientific community. He and his other colleagues (Sorensen et al. 2010) applied co-authorship network analysis to perform an eigen decomposition of the Medical Subject Heading (MeSH) terms from AD literature. Recently, Chen et al. (2014) investigated the trend of AD research and the order of most tolerated or effective drugs for AD treatment, focusing on the publication of cholinesterase inhibitor research. However, none of the previous studies conduct the comprehensive analysis both at the macro level (year, journal or proceeding, author, etc.) and at the micro level (bio-entity) at the same time to broadly examine the landscape of AD literature.

Specifically, in order to perform analysis at a micro-level, we apply the concept of co-occurrence to the AD literature, stemmed from text mining techniques. Text mining refers to the extraction of hidden and useful information or knowledge by processing unstructured text data (Erhardt et al. 2006). It has been abundantly applied to biomedical text since the late 1990s, and the field of AD research was no exception. At first, Smalheiser and Swanson developed ARROWSMITH which helps researchers generate and test highly possible hypotheses linking AD and estrogen (Smalheiser and Swanson 1996) or indomethacin (Smalheiser and Swanson 1998), based on text mining on titles of Medline articles. The recent research (Li et al. 2009), which created AD-specific drug-protein connectivity maps based on literature mining on PubMed abstracts, also used the maps for obtaining nontrivial knowledge about related genes, proteins, candidate drugs and protein therapeutic/toxicological profiles of the candidate drugs as well as producing a new hypothesis. Similarly, Krauthammer et al. (2004) succeeded in figuring out genes which make people vulnerable to AD by parsing full-text scientific articles and analyzing species-specific molecular networks. Al-Mubaid and Singh (2005) derived the significant associations between AD and specific proteins by analyzing Medline abstracts with the concepts of expectation (ex), evidence (ev), and Z-scores. These studies have enabled researchers to better understand the multifactorial AD and even contributed to the improvement of AD treatment. Unlike previous studies utilizing text mining for the AD literature, our study provides the unique landscape of the AD literature, focusing on the relationships of bio-entities and the distribution of main topics to help researchers to recapitulate the AD research field.

Proposed approach

In this section, we describe document representation techniques used in text mining and explain how we construct graph representations of a text document, starting from a small set of concepts and expanding it into a rich graph with additional semantic information. The discussion also explains the motivation behind using such representations for extraction tasks.

Concept graphs

Concept graphs, which consist of sets of nodes and edges, represent the text documents. We start with a small set of concept nodes extracted from a document's meta-data and add external concept nodes with the corresponding relationships (edges).

As illustrated in Fig. 1, the proposed representation is constructed using a small set of document features and expanded into a richer representation using domain concepts and semantic relationships. In this representation we don't consider any statistical information derived from the text which makes the proposed method less dependent on a document's content. In addition, using external domain knowledge, the representation is projected into a more domain-specific feature space. Starting with a handful set of entities representing a document and mapping those into predefined concepts and relations, we represent each document using a graph where nodes represent concepts that might or might not appear in the text, and edges represent semantic relationships that exist among the concepts in a certain domain.

The dataset used in the experiments is a collection of articles collected from medical journals. In addition, we use UMLS as an external knowledge base of biomedical concepts. UMLS provides a comprehensive vocabulary database and ontology of biomedical concepts and relationships among them. UMLS is maintained by the National Library of Medicine[2] and is updated regularly. Currently, it contains records of over 1 million concepts with different naming conventions and collected from different medical and health related sources.

For each article in the dataset, we extract biological entities by a named entity recognition technique. Those are then mapped into predefined UMLS concepts, which we refer to as key concepts. In the mapping process, we attempt to find either a first-best (fb) match or n-gram (ng) matches of a keyphrase into a UMLS concept. For instance, if the phrase 'Atypical antipsychotic drugs' is extracted from either title or abstract, it would be mapped to the concept 'Antipsychotic Drugs' using first-best mapping since 'Antipsychotic Drugs' is the first successful match with a maximum length (number of terms) and since the whole initial phrase doesn't exist in UMLS. Using n-gram mapping, it would be mapped to all

[2] "National Library of Medicine—National Institutes of Health" [Online]. http://www.nlm.nih.gov/. Accessed 8 Feb 2013.

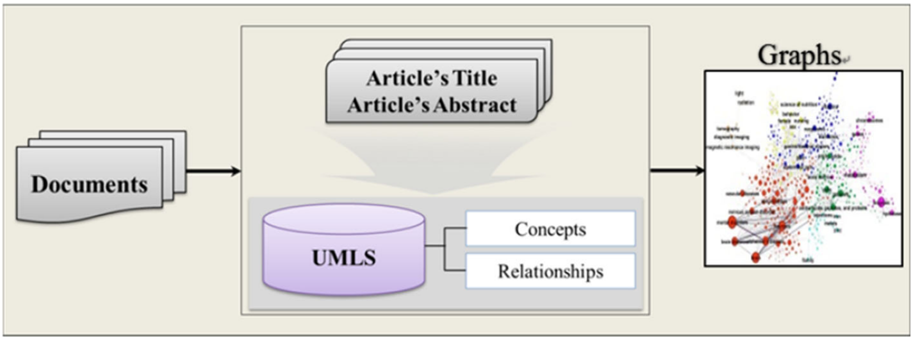

Fig. 1 From documents into graphs

combinations of concepts that correspond to the terms in the phrase and exist in UMLS, in this case: 'Antipsychotic Drugs', 'Antipsychotic', and 'Drugs'. Figure 1 above shows an illustration of the graph construction process.

After extracted entities are mapped into unique UMLS concepts, the obtained list is used as the base nodes list of the concept graph. The graph is then expanded by adding related concepts queried from UMLS. Relationships are available as pairs of related concepts and semantic relationships between them. Examples of related concepts in UMLS are: 'Anxiety—mental disorders' and 'Pathologic Process—psychological stress'. The semantic relationships are typically synonym, parent–child, sibling, broad, and narrow relationships. The related concepts are added to the graph as new nodes where the relationships are represented by edges that connect those nodes. Upon adding new nodes, if a concept is related to an existing concept in the graph, an edge is also added to link them together. This process is also parameterized as we choose a variable level of related concepts to be added to the graph. In the experiments we construct graphs of one level of related concepts as well as two levels, where concepts related to the related concepts are also added. This is meant to increase the degree of the graph representation by adding more domain knowledge that could be more discriminative with respect to a document's class. Adding more levels of related nodes however, would increase the degrees of graphs exponentially and could add some noisy and irrelevant concepts to the representation. For that reason we expand the graphs up to two levels at most. Figure 2 shows an example of concept nodes and the relationship edges that connect them.

Topic modeling

Topic models are based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated.

Document modeling in text mining is a technique that expresses an individual document and the collection of documents by term appearing in documents. Topic modeling is one of the document modeling techniques, and LDA, standing for Latent Dirichlet Allocation proposed by Blei et al. (2003), is one of the earliest topic modeling techniques that is based on a graph model with an assumption of Dirichlet prior-based topic distribution. In other words, LDA represents documents as mixtures of topics that spit out words with certain probabilities. The topic modeling technique used in this paper is Dirichlet-multinomial

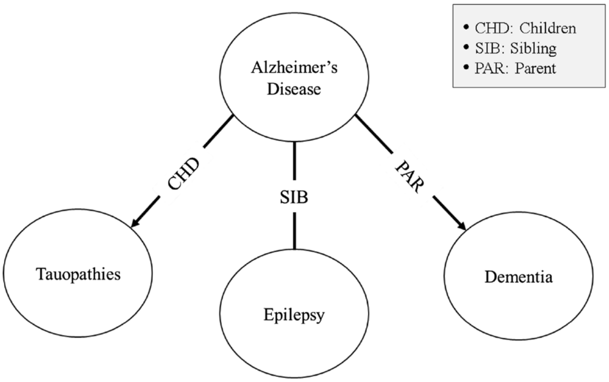

Fig. 2 Concept nodes and relationships

regression (DMR) proposed by Mimno and McCallum (2008). DMR is an extension of Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003), and allows conditioning on arbitrary document features by including a long-linear prior on document-topic distributions that is a function of the features of the document such as author, publication venue, references, and dates. For each document d, let x_d be a feature vector representing metadata. Given the prior distribution of $N(0, \Sigma)$.and hyper-parameters β, the generative process for documents and their words is as follows:

(1) For each topic t, draw $\varnothing_{\mathrm{t}}$ $\sim$Dir(β) noting that Dir(β) is a distinct Dirichlet distribution with the Dirichlet prior on the topic-word distribution (a.k.a. hyper-parameters), β.
(2) For each document d, draw $\theta_{\mathrm{d}} \sim \mathrm{Dir}(\alpha_{\mathrm{d}}) = \mathrm{Dir}(\exp(\tau_{\mathrm{d}}))$with $\tau_{\mathrm{d}} \in \tau$ noting that a per-document α_{d}, the parameters of a Dirichlet distribution and τ_{d} is a covariance function $f(y_{\mathrm{d}}, x_{\mathrm{k}})$ where y_d is the observed attribute vector of document d and x_k is a vector of metadata.
(3) For each word w,
   - Draw $Z_{\mathrm{d,w}}$ $\sim$Multi(θ_d) noting that $Z_{\mathrm{d,w}}$ is topic assignment of a word t_{w} and θ_{d} is topic proportion of a document d.
   - Draw $T_{\mathrm{d,w}}$ $\sim$Multi($\emptyset_{Z_{d,w}}$) noting that $T_{d,w}$ is w-th word of a document d and $\varnothing_{\mathrm{t}}$ is preference of a topic t over the vocabulary with $\sum_n \emptyset_{t,n} = 1$.

For DMR topic modeling, we set three fixed parameters: σ^2, the variance of the prior on parameter values for prior distribution; β, the Dirichlet prior on the topic-word distributions; and |T|, the number of topics. In this study, we hire temporal LDA to compare and analyze relationship between topics extracted from scientific publications.

Results and discussion

In this section, we provide the detailed description of data collection (Table 1) and experimental results followed by comprehensive discussion of the results. We collected 96,279 articles over the period of pre-1950 to 2014, retrieved from PubMed by the query [“Alzheimer disease” OR “Alzheimer’s disease”]. Among them, 96,081 articles are the subject of analysis as they have both titles and abstracts which allow for content analysis. The analysis of the results is conducted in three different ways: productivity analysis, network analysis, and content analysis.

Productivity analysis

Productivity analysis is conducted from four angles (Year, Journal/Proceeding, Author, and MeSH Term) to better understand the productivity of literature in the research field of AD.

Productive years

Figure 3 shows the productivity based on the published year. It is evident that the productivity of AD research has increased exponentially. It thus can be inferred that the research field would grow further in the near future. Table 2 describes time-based productivity focusing on the 2000s. The most productive year was the year 2013. The number of papers in 2013 accounts for 10.59 % of 69,254 articles in the 2000s, and about 7.64 % of all 96,081 articles in the analysis. The productivity of the year 2014 has not been completely calculated since we collected data on April, 2014.

Productive journals/proceedings

Table 3 presents each journal's productivity. Among 4,480 journals, Journal of Alzheimer's Disease (JAD) is the most productive journals in AD research. It is understood that

Table 1 Data collection

Number of retrieved papers	96,279
Number of papers in the analysis	96,081
Period	pre-1950–2014
Number of Journals	4,480
Number of Authors	253,575
Number of MeSH terms	16,176

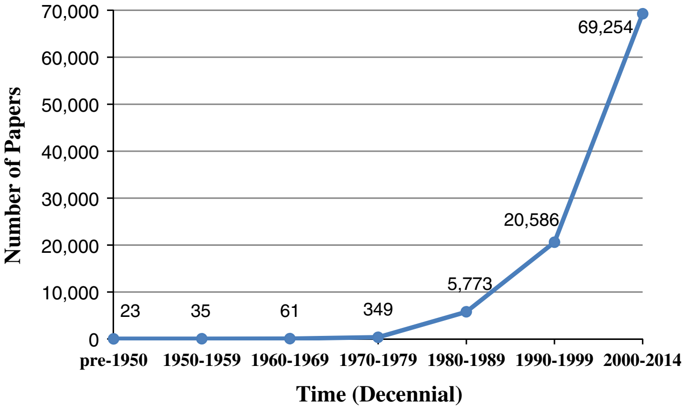

Fig. 3 Time-based productivity for all of the periods (decennial)

Table 2 Time-based productivity from 2000 to 2014 (yearly)

Year	Number of Papers	Ratio (%)	Ranking
2000	3,968	5.73	11
2001	5,280	7.62	6
2002	5,659	8.17	5
2003	3,605	5.21	12
2004	1,512	2.18	15
2005	1,733	2.50	14
2006	4,174	6.03	10
2007	4,367	6.31	9
2008	4,686	6.77	8
2009	5,123	7.40	7
2010	5,755	8.31	4
2011	6,028	8.70	3
2012	6,779	9.79	2
2013	7,336	10.59	1
2014	3,249	4.69	13
	69,254	100.00	

Table 3 Journal's productivity

Ranking	Journal Name	Number of papers	Ratio (%)
1	Journal of Alzheimer's Disease: JAD	2,784	2.90
2	Neurobiology of Aging	2,432	2.53
3	Neurology	2,036	2.12
4	Neuroscience Letters	1,755	1.83
5	The Journal of Biological Chemistry	1,377	1.43
6	Journal of Neurochemistry	1,263	1.31
7	Alzheimer Disease and Associated Disorders	1,231	1.28
8	Brain Research	1,216	1.27
9	Dementia and Geriatric Cognitive Disorders	1,166	1.21
10	PloS One	1,128	1.17

top three journals including JAD, Neurobiology of Aging, and Neurology are fairly influential in the corresponding research field, considering the sum of their ratios (7.55 % of 96,081 papers). Meanwhile, the most productive proceeding, Proceedings of the National Academy of Sciences of the United States of America, is ranked as 13th by having 891 AD papers. As indicated by top ten ranked journals and proceedings, journals are certainly more dominant for publishing AD related articles than proceedings.

Productive authors

Table 4 demonstrates top ten authors with their affiliation, who have published a number of papers in AD research, among a total of 253,575 authors. To disambiguate author names,

Table 4 Author's productivity

Ranking	Author Name	Affiliation	Number of Papers	Ratio (%)
1	Mark A. Smith	Department of Pathology, Case Western Reserve University, USA	309	0.322
2	George G. Perry	Department of Pathology, Case Western Reserve University, USA	298	0.310
3	Kaj K. Blennow	Department of Clinical Neuroscience, Gothenburg University, Sweden	290	0.302
4	Bengt.B. Winblad	Department of Clinical Neuroscience, Huddinge University Hospital, Sweden	272	0.283
5	David A. Bennett	Rush Alzheimer's Disease Center, Rush University Medical Center, USA	257	0.267
6	John Q. Trojanowski	Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, USA	235	0.245
7	Philip P. Scheltens	Department of Neurology and Alzheimer Centre, Vrije Universiteit Medical Centre, Netherlands	223	0.232
8	Bengt B. Winblad	Department of Neurotechnology, Huddinge University Hospital, Sweden	222	0.231
9	Konrad K. Beyreuther	ZMBH, Center for Molecular Biology, University of Hei delberg, Germany	221	0.230
9	John C. Morris	Department of Neurology, Washington University School of Medicine, USA	221	0.230

we first sort out author names with their affiliations automatically extracted from the PubMed records. We then manually combine the papers written by an identical person. All of top ten authors have written more than 220 papers. It is discovered that half (5) of top authors are from United States of America. The other countries with active AD research are Sweden (3), Netherlands (1), and Germany (1), based on the affiliation of the authors. Top two authors have an affiliation of Case Western Reserve University (CWRU) in USA. In addition to being a professor at CWRU, Mark Smith (1st) served as the Director of Basic Science Research at the University Memory and Aging Center, Editor-in Chief of JAD, and the member of the Editorial Board of over 20 leading journals. George Perry (2nd), who holds an adjunct professional appointment at CWRU, is now Dean of the College of Sciences and Professor of Biology at the University of Texas at San Antonio. Both of them are well recognized in the field of AD research. Kaj Blennow (3rd), who is a professor at Gothenburg University and also Senior Consultant at Neurochemistry Laboratory of Sahlgren's University Hospital in Sweden. His major research interests are cerebrospinal fluid biochemical markers for the clinical diagnosis of AD, genetic mechanisms of AD, or neurochemical pathogenesis of AD, and he wrote more than 380 research papers in peer-reviewed journals. What top ten authors have in common is that they are productive researchers who have actively involved in AD research and have a high authority in the related research center or laboratory.

Frequent MeSH terms

Table 5 shows top 20 MeSH terms assigned to 96,081 papers. In Table 5, ratio means which percentage of those 96,081 papers contain the MeSH term. Overall, top MeSH terms reflect main subjects in the research field of AD. Not surprisingly, a high percentage of papers have *Humans* and *Alzheimer disease* as their MeSH terms. We can see terms closely related to AD such as *aged/aging, peptides, dementia, mice/rats*, etc.

Network analysis

The global concept graph, which is an integration of 96,081 individual concept graphs from 96,081 articles, provides biological entities involved in AD research and the relationships among them. It is visualized using Gephi, an open source for the network analysis and visualization (Bastian et al. 2009). The global graph consists of 20,409 vertexes and 50,312 edges at the initial stage. Each vertex refers to a biological entity derived from articles, and matches with unique CUI. The edge means the relationship between entities, which is defined by UMLS. The edge's weight represents the co-occurrence frequency of two entities in a specific sentence of articles.

For better visualization, we removed the edges with weights under 300 and separate nodes that have no links to other nodes. We further deleted 24 nodes among top degree centrality nodes since they are too common words to have connection with AD. The examples include taxonomic, historical aspects qualifier, anatomy & histology (qualifier), diseases (mesh category), publication characteristics, named groups (mesh category) and equipment (mesh category). The final concept graph ends up with 3,634 nodes and 5,538 edges. We visualize the graph with some nodes and their labels expressed for each community after applying community detection algorithm as shown in Fig. 4. The final global concept graph is examined with the following three methods: co-occurrence frequency analysis, centrality analysis, and community detection.

Table 5 MeSH term's frequency

Ranking	MeSH Term	Number of papers	Ratio (%)	Ranking	MeSH Term	Number of papers	Ratio (%)
1	Humans	73,021	76.00	11	Mice	10,011	10.42
2	Alzheimer disease	61,978	64.51	12	Dementia	8,915	9.28
3	Aged	35,617	37.07	13	Neuropsycho-logical tests	8,039	8.37
4	Male	33,373	34.73	14	Neurons	7,833	8.15
5	Female	31,291	32.57	15	Rats	7,281	7.58
6	Animals	25,616	26.66	16	Cognition disorders	7,135	7.43
7	Middle Aged	18,966	19.74	17	Adult	6,872	7.15
8	Aged, 80 and over	17,449	18.16	18	Amyloid beta-protein precursor	6,411	6.67
9	Brain	16,455	17.13	19	Aging	6,092	6.34
10	Amyloid beta-Peptides	14,211	14.79	20	Peptide fragments	5,873	6.11

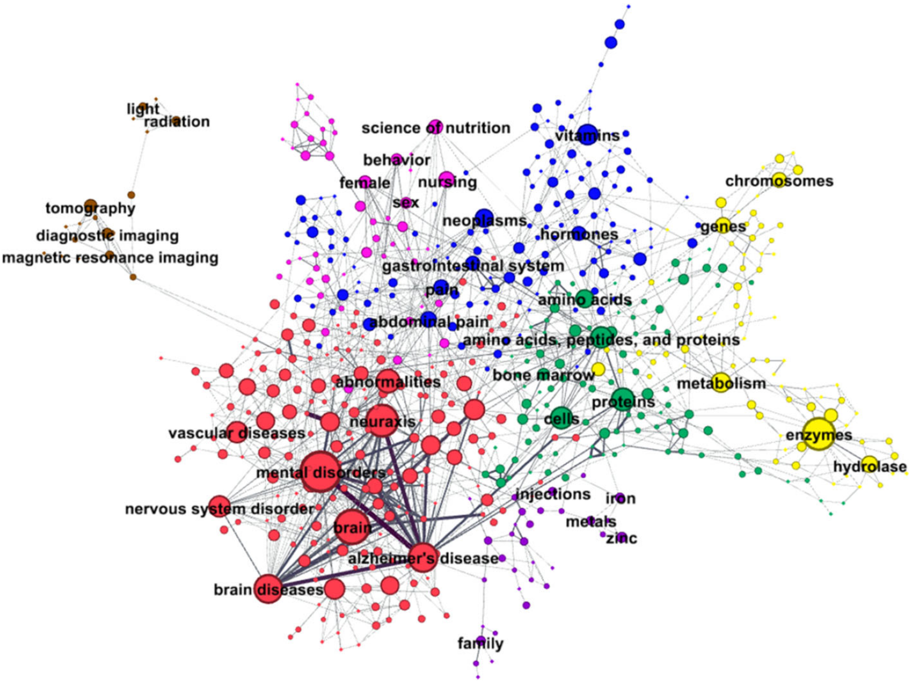

Fig. 4 Visualization of Alzheimer's disease literature

Co-occurrence frequency analysis

To understand what are the major biological entities and their relationships in AD research, we first construct the bio-entity pairs in order of the highest co-occurrence frequency. Table 6 shows top ten bio-entity pairs and their relations that have high co-occurrence frequency. RO and CHD are frequently seen UMLS relationship types as each of them appears three times among top ten ranks. RO refers to any relationships other than synonymous, narrower, or broader and CHD is the child relationship in a Meta-thesaurus source vocabulary.

According to Table 6, the entities *Alzheimer's disease*, *mental disorders*, and *disease* appear many times, which is inevitable since the literature is about AD. As the entity *Alzheimer's disease* is linked with *mental disorders*, *neuraxis*, and *brain diseases*, we can infer that AD is deeply related to brain. The highest frequency pair is the one between *glycogen storage disease* and *glycogen storage disease type vi* while they are similar or alike to each other as RL indicates. The entity *glycogen storage disease type vi* is also frequently connected to the other entity *liver*, with different relationships from synonymous, narrower, or broader.

Centrality analysis

To recognize the core biological entities in the literature of AD research, we analyze top ten centrality nodes by five well-known centrality measures: degree centrality, weighted degree centrality, closeness centrality, betweenness centrality and PageRank. The details of the centrality measures are offered by Wasserman and Faust (1994) and Brin and Page (1998).

Degree centrality. Degree centrality of a specific node means the number of edges that is connected to that node (Wasserman and Faust 1994). Table 7 describes top ten bio-entities ranked by degree centrality. The average degree centrality value is 3.048. The bio-entity with the highest value of degree centrality is *mental disorders*. All the entities except the *nervous system disorder* which is ranked as 10th are also included in top ten bio-entities rankings of other centrality measures. The entity *enzymes* tied with *Alzheimer's disease* for 3rd place seems to appear with high degree centrality value because some of enzymes are known as causes of AD (e.g. caspase-3 enzymes) while others are applied to cure the disease. Several proteins are also widely known causes of AD (e.g. Beta-amyloid Protein) and can be seen in the ranking of MeSH term's frequency (Table 5), which makes the entity *proteins* ranked as 8th.

Weighted degree centrality. Weighted degree centrality is an extended version of degree centrality, calculated by summing the frequency of every node pair for a specific node (Wasserman and Faust 1994). Table 8 shows top ten bio-entities according to weighted degree centrality. The value of weighted degree centrality on average is 6,640.683. We notice that six of core bio-entities in Table 8 are among those of top degree centrality.

Table 6 Co-occurrence frequency and relationship between two bio-entities (top ten)

Ranking	Frequency	Bio-entity pair		Relation
1	55,909	Glycogen storage disease	Glycogen storage disease type vi	RL
2	55,247	Glycogen storage disease type vi	Liver	RO
3	54,089	Mental disorders	Alzheimer's disease	PAR
4	50,661	Behavior disorders	Mental disorders	RN
5	50,621	Behavior disorders	Disease	CHD
6	50,620	Abnormalities, radiation-induced	Disease	CHD
7	48,169	Acute disease	Disease	CHD
8	44,595	Alzheimer's disease	Neuraxis	RO
9	41,469	Alzheimer's disease	brain diseases	RO
10	34,937	Mental disorders	dementia	RQ

Table 7 Top ten bio-entities by degree centrality

Ranking	Bio-entity	Degree centrality
1	Mental disorders	51
2	Brain	44
3	Neuraxis	41
3	Enzymes	41
5	Alzheimer's disease	37
6	Brain diseases	36
7	Abnormalities	30
8	Proteins	29
9	Cells	28
10	Nervous system disorder	27

Table 8 Top ten bio-entities by weighted degree centrality

Ranking	Bio-entity	Weighted degree centrality
1	Brain	447,300
2	Alzheimer's disease	419,866
3	Dementia	336,506
4	Mental disorders	312,874
5	Brain diseases	246,234
6	Disease	235,305
7	Neuraxis	211,577
8	Proteins	185,231
9	Patients	172,103
10	Neurons	163,549

Table 9 Top ten bio-entities by closeness centrality

Ranking	Bio-entity	Closeness centrality
1	Transduction, genetic	12.4,560
1	Transformation, bacterial	12.4560
3	Genes, transgenic, suicide	11.4576
3	Gene transfer techniques	11.4576
5	Transfection	11.4563
6	Health manpower	11.0054
7	Type c phospholipases	10.8995
7	Alpha toxin, clostridium perfringens	10.8995
7	Phosphoric diester hydrolase	10.8995
10	European continental ancestry group	10.5855

However, entities such as *dementia*, *disease*, *patients*, and *neurons* are unique in the ranking of the weighted degree centrality. With degree value weighted, core entities become much more like those in the frequent MeSH term list (Table 5).

Closeness centrality. Closeness centrality means the inverse of the sum of total distances from a particular node to every other node in the network, eventually focusing on the node's extensibility over the network (Wasserman and Faust 1994). Table 9 represents top ten bio-entities based on closeness centrality. The average closeness centrality value is 6.194. The ranking of the closeness centrality involves many unique bio-entities such as *health manpower*, *type c phospholipases* and *alpha toxin, clostridium perfringens*, compared to the rankings of other centrality measures. This is mainly due to the formula of closeness centrality. The formula enables entities with low co-occurrence frequency to have a smaller sum of total distances and accordingly bigger centrality value. In fact, among the top ten bio-entities, we discover some experimental terms like *transduction, transformation*, and *transfection* as well as gene-related technological terms like *gene, transgenic, suicide* and *gene transfer techniques*.

Betweenness centrality. Betweenness centrality is defined as the number of shortest paths passing through a given node (Wasserman and Faust 1994). Table 10 presents top ten bio-entities on the basis of betweenness centrality. The average centrality value is

8,051.002. We detect bio-entities which are common among top-entities of other centrality measures. The examples are *alzhiemer's disease*, *proteins*, *brain, mental disorders,* and *cells*. The entity of the highest betweenness centrality value is *amino acids, peptides, and proteins*, and its value is about 1.75 times more than the value of the 2nd top bio-entity. This indicates that the 1st bio-entity plays a significant role in mediating the relations between specific nodes in the network. In fact, it has a profound connection with the study for the cause of AD.

PageRank. PageRank estimates the importance of a particular node based on the sum of the ranks of the number of its incoming links (Brin and Page 1998). Table 11 indicates top ten bio-entities according to PageRank. The average value of PageRank is .00027518. The top bio-entities of PageRank resemble those of degree centrality (8 out of 10), with a change in the rank order. However, the ranking by PageRank still shows unique bio-entities like *metabolism.*

Community detection

We perform community detection with a modularity algorithm (Blondel et al. 2008) after leaving giant components and setting the resolution value as 5 considering the huge size of

Table 10 Top ten bio-entities by betweenness centrality

Ranking	Bio-entity	Betweenness centrality
1	Amino acids, peptides, and proteins	1,085,667.40
2	Alzheimer's disease	621,156.23
3	Proteins	472,137.36
4	Brain	435,992.11
5	Mental disorders	342,662.27
6	Diagnosis	309,762.28
7	Cells	282,360.40
8	Genes	280,848.15
9	Body part	270,838.01
10	Acquired immunodeficiency syndrome	261,028.44

Table 11 Top ten bio-entities by PageRank

Ranking	Bio-entity	Page rank
1	Mental disorders	0.00272686
2	Enzymes	0.00267630
3	Brain	0.00251523
4	Neuraxis	0.00212942
5	Proteins	0.00201112
6	Alzheimer's disease	0.00196997
7	Abnormalities	0.00184893
8	Amino acids	0.00184885
9	Metabolism	0.00181797
10	Brain diseases	0.00180405

the final concept graph (Lambiotte et al. 2009). The number of automatically grouped communities by the modularity algorithm turns out to be 7 as in Fig. 4. Two grand communities which account for almost half of the network together are red and blue ones. The largest community in red (32 %) contains *mental disorders, Alzheimer's disease, brain, neuraxis* and *nervous system disorder*, which are associated with brain disease in general. Many entities with high centrality values belong to this red community. The next largest community's (22 %) representative nodes (blue) are *vitamins, hormones, pain,* and *neoplasms*, which are related to (mostly negative) change in the body. The 3rd biggest community (green, 17 %) has cells or proteins including *amino acids, peptides, and proteins* which ranked as 1st in the ranking by betweenness centrality, mediating the connection between nodes. The community in yellow (13 %) consists of *enzymes* and *metabolism*, pink one (8 %) of *nursing* and *nutrition* for the cure of disease, purple one (5 %) of some metallic elements, and finally brown one (3 %) with entities related to the brain image for the diagnosis of AD. Overall, each community consists of bio-entities closely related to each other inside the community with having a specific subject or topic in AD research field.

Content analysis

DMR-based topic modeling identifies 16 major topics in the field of AD research across 96,081 articles. Table 12 represents labeled topics and 8 keywords within each topic.

The first three topics (AD-associated Factor, Transgenic Mouse, and Dementia) focus on general or broad study of AD. As seen in the keywords of Topic 1, some researchers have tried to find risk factors of AD and verify age or gender is one of them. In recent years, it was found that older age is considered to be a main risk factor (Hebert et al. 2001; Seshadri et al. 1997) while gender is not (Bachman et al. 1993; Barnes et al. 2003; Evans et al. 2003; Hebert et al. 2001; Kukull et al. 2002; Miech et al. 2002; Rocca et al. 1998). The keywords of Topic 2 show that transgenesis is a core technique for AD research. Those of Topic 3 include several diseases closely related to AD. Next five topics (APP/PS, Beta-amyloid Protein, Tau Protein, Oxidative Stress/Mitochondria, and ApoE) are associated with widely known causes or related bio-entities of AD. Topic 9 (CSF Levels) and Topic 10 (Brain Image) are about the diagnosis and the following four topics (Brain Plaques/Tangles, Cholinergic System, Induced Cell Activity and Memory/Cognitive Impairment) have to do with symptom. The remaining two topics (Treatment and Caregiving) are for AD treatment. Topic 15 relates to various drugs (e.g. memantine, donepezil, and rivastigmine) while Topic 16 refers to caregiving for AD patients which can be a big burden for their families. These topic modeling results enable us to grasp prevalent topics in the literature of AD.

Each topic's relative distribution over time is described in Fig. 5. We examined the topic distribution for the period of 2000 to 2013, considering a sufficient data size. Overall, there is a widening gap of the topic distribution over time while all topics have a similar distribution early in 2000. All the topics can be clearly grouped into three types by trend. First, three topics which are Transgenic Mouse, Tau Protein and Brain Image have a rising trend (marked with triangles). Transgenic Mouse especially shows the sharpest growth through time with a notably high distribution ratio near the year 2013. The growth of the other two topics seems relatively gradual. It can be inferred that AD researchers nowadays tend to understand the disease and solve the treatment problem through advanced technologies. Two topics which are Dementia and Brain Plaques/Tangles, on the other hand,

Table 12 DMR-based topic modeling results

Topic 1		Topic 2		Topic 3		Topic 4	
AD-associated factor		Transgenic mouse		Dementia		APP/PS*	
Risk	Factors	Mice	Brain	dementia	pd*	app*	Precursor
Age	Women	Transgenic	Amyloid	Disease	Clinical	Amyloid	Secretase
Years	Older	Memory	Levels	Patients	Syndrome	Protein	Bace
Associated	Elderly	Model	Effects	dlb*	ftd*	Beta	Presenilin
Topic 5		Topic 6		Topic 7		Topic 8	
Beta-amyloid protein		Tau protein		Oxidative stress/mitochondria		ApoE	
Amyloid	apos*	tau	gsk*	stress	Antioxidant	apoe	Apolipoprotein
Beta	Protein	Protein	Neurofibrillary	Mitochondrial	Activity	Gene	Association
Peptide	Aggregation	Phosphorylation	phf*	Brain	Levels	Allele	Genotype
Formation	Binding	Kinase	Tangles	Iron	Oxygen	Epsilon	Polymorphism
Topic 9		Topic 10		Topic 11		Topic 12	
CSF levels		Brain image		Brain plaques/tangles		Cholinergic system	
ad	apos	Brain	Hippocampal	Brain	Neurofibrillary	Receptor	Inhibitors
Patients	Plasma	Imaging	Cortex	Plaques	Cerebral	Activity	Acetylcholineste-rase
Levels	Cholesterol	mri*	Tomography	Neurons	Tangles	Cholinergic	Acetylcholine
csf*	Higher	Temporal	Regional	Amyloid	Deposits	Binding	Effects
Topic 13		Topic 14		Topic 15		Topic 16	
Induced cell activity		Memory/cognitive impairment		Treatment		Caregiving	
Induced	Expression	Memory	Performance	Treatment	Memantine	Care	Nursing
Cell	Death	Patients	Cognitive	Placebo	Drug	Caregivers	Burden

Table 12 continued

Topic 13		Topic 14		Topic 15		Topic 16	
Induced cell activity		Memory/cognitive impairment		Treatment		Caregiving	
Activation	Microglia	Subjects	Task	Donepezil	Dose	Health	Social
Neurons	Apoptosis	Impairment	Deficits	Rivastigmine	Effect	Family	Costs

*apos** apolipoproteins, *app** amyloid precursor protein, *csf** cerebrospinal fluid, *dlb** dementia with lewy bodies, *ftd** frontotemporal dementia, *gsk** glycogen synthase kinase, *mri** magnetic resonance imaging, *pd** Parkinson's Disease, *ps** Presenilin, *phf** paired helical filaments

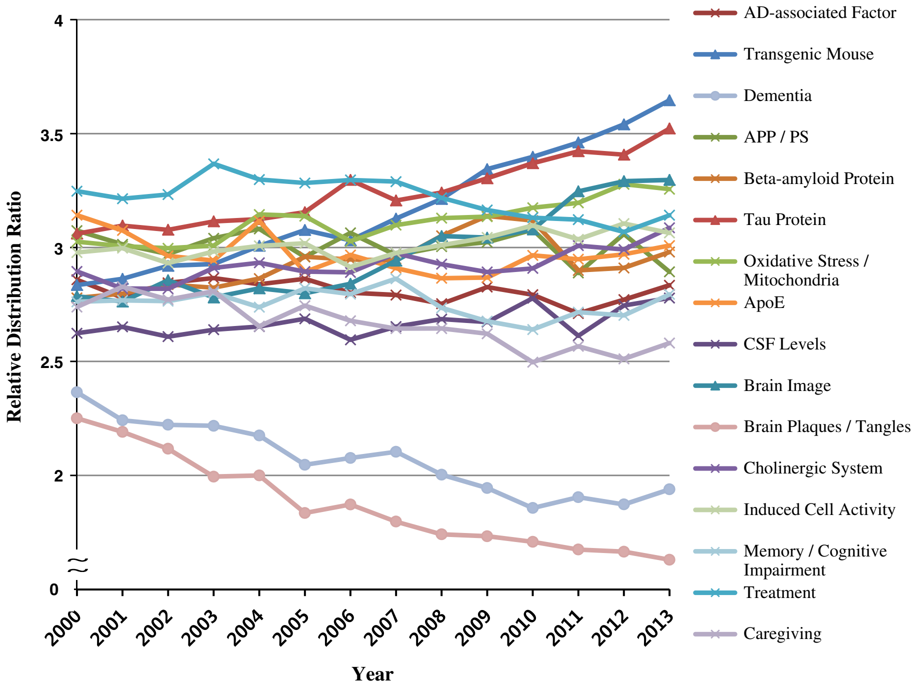

Fig. 5 Time series topic distribution from the year 2000–2013

have a downward trend (marked with circles). The rest of topics maintain a monotonous tendency, sometimes with a little fluctuation (marked with X).

Discussion

The present study shows certain aspects of AD research consistent with the previous studies with some distinct properties. In particular, in productivity analysis, we confirm that the most productive journal/country is similar to the research of Ansari et al. (2006), and the productive AD researchers, and frequent MeSH terms are similar to the study of Sorensen (2009). But there are still subtle differences. For instance, we observe the reversed rank order of 2nd and 3rd productive journal compared to the article of Ansari et al. (2006). Another example includes that Mark Smith who is considered as the most productive AD author here took the 3rd place in the Sorensen's paper (2009). Several differences are due to discrepancy in datasets, but the results from this paper hold more recency.

Network analysis reveals that most frequently mentioned entities in the AD literature are associated with the semantic type "disease or syndrome" to which entities like *Alzheimer's disease, brain disease* and *glycogen storage disease* belong, followed by "Body System" with entities *brain* and *neuraxis*. The results do not show novel relationships but rather affirm core entities and the relationships among them frequently studied in the literature. For instance, the top ranked entity *Alzheimer's disease* is linked to brain related entities. One interesting observation is that the relationship between the entities *glycogen* *storage disease* and *glycogen storage disease type vi*, whose semantic types are “disease or syndrome” and relation type is RL (meaning “similar”) occurs many times.

Topic model identifies several core research areas of AD. The results show that AD pathogenesis-related entities such as Beta-amyloid Protein and Tau Protein are major topics in the AD literature. In trend analysis, topics showing the rising trend reflect recent rigorous research efforts on applying genomics for studying AD (Ravetti et al. 2010; Orešič et al. 2010; Thota et al. 2007). These trends confirm that the combination and coordination of the advanced technology with AD research deepens the understanding of AD and enriches the corresponding research field. The topic showing a clear falling trend is related to dementia and brain plaques/tangles. This topical trend may be attributed to a higher level of concentration to AD itself, rather than a group of brain diseases, and a lower level of interest in the relationship between brain plaques/tangles and AD since this relationship was discovered decades ago.

Conclusion

Alzheimer’s disease is a genetic disease that affects the function of the brain and causes brain degeneration over time. It is the most common form of progressive dementia in elderly people, and become a more severe international health issue in the near future (which is predicted to affect 1 in 85 people globally by 2050 by Brookmeyer et al. 2007).

In the present study, we attempted to identify the current landscape of AD research. To this end, we retrieved 96,279 articles with the query term [“Alzheimer disease” OR “Alzheimer’s disease”] and used 96,081 out of them, which have titles and abstracts from PubMed. Unlike previous studies driven by bibliometrics, we utilized concept graphs and topic modeling to map out the field of AD. By applying network and content analysis to the results of concept graph and topic modeling, we were able to identify major biological entities mentioned in the AD literature and the salient relationships among entities. In addition, we enabled to discover major topics studied in the literature and topic shift over time.

As a result, the productivity analysis reveals the dynamics of AD research led by several famous experts especially in the last ten years while the year 2013 is found to be the most productive. In the network analysis, top bio-entities are partly similar to frequently-appeared MeSH terms, however, they are more concentrated in specific topics (e.g. diagnostic imaging, neoplasm, etc.) than MeSH terms. Content analysis identifies salient topics associated with the cause of AD and the recent rising trend of topics in the field of the advanced science and technology, such as transgenics and brain imaging techniques.

The major limitation of the study is that data collection is limited to PubMed only. The composite dataset may provide a clearer picture of the field. Thus, as the follow-up study, we plan to expand the dataset by collecting data from Web of Science or using full-texts from PubMed Central. In addition, we will attempt to discover previously unknown relations among biological entities in the AD literature with hope of finding a clue of missing puzzle to cure AD.

Acknowledgments This work was supported by the Bio-Synergy Research Project (2013M3A9C4078138) of the Ministry of Science, ICT and Future Planning through the National Research Foundation.

References

Al-Mubaid, H., & Singh, R. K. (2005). A new text mining approach for finding protein-to-disease associations. *American Journal of Biochemistry and Biotechnology, 1*(3), 145.

Andreasen, T., Bulskov, H., Jensen, P. A., & Lassen, T. (2009). Conceptual indexing of text using ontologies and lexical resources. Presented at *the Proceedings of the eighth international conference on flexible query answering systems* (Vol. 5822, pp. 323–332). Berlin: Springer.

Ansari, M. A., Gul, S., & Yaseen, M. (2006). Alzheimer's disease: A bibliometric study. *Trends in Information Management (TRIM)*, *2*(2), 130–140.

Bachman, D., Wolf, P. A., Linn, R., Knoefel, J., Cobb, J., Belanger, A., … D'Agostino, R. (1993). Incidence of dementia and probable Alzheimer's disease in a general population The Framingham Study. *Neurology*, *43*(3 Part 1), 515–515.

Barnes, L., Wilson, R., Schneider, J., Bienias, J., Evans, D., & Bennett, D. (2003). Gender, cognitive decline, and risk of AD in older persons. *Neurology, 60*(11), 1777–1781.

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks*. (pp. 361–362). Presented at the International AAAI Conference on Weblogs and Social Media, ICWSM 2009.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bleik, S., Song, M., Smalter, A., Huan, J., & Lushington, G. (2009). *CGM: A biomedical text categorization approach using concept graph mining* (pp. 38–43). Presented at the IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2009, BIBMW 2009.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1), 107–117.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia, 3*(3), 186–191.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. *Proceedings of 3rd annual symposium on document analysis and information retrieval*, *48113*(2), 161–175.

Chen, H., Wan, Y., Jiang, S., & Cheng, Y. (2014). Alzheimer's disease research in the future: bibliometric analysis of cholinesterase inhibitors from 1993 to 2012. *Scientometrics, 98*(3), 1865–1877.

Chen, Y.-M., Wang, X.-L., & Liu, B.-Q. (2005). Multi-document summarization based on lexical chains. 2005. Presented at the *Proceedings of 2005 IEEE international conference on machine learning and cybernetics* (Vol. 3, pp. 1937–1942).

Damashek, M. (1995). Gauging similarity with *n*-grams: Language-independent categorization of text. *Science, 267*(5199), 843–848.

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing and Management, 43*(6), 1705–1714.

Erhardt, R. A., Schneider, R., & Blaschke, C. (2006). Status of text-mining techniques applied to biomedical text. *Drug Discovery Today, 11*(7), 315–325.

Evans, D. A., Bennett, D. A., Wilson, R. S., Bienias, J. L., Morris, M. C., Scherr, P. A., et al. (2003). Incidence of Alzheimer disease in a biracial urban community: Relation to apolipoprotein E allele status. *Archives of Neurology, 60*(2), 185–189.

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. Presented at the *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 19–25). New York: ACM.

Hebert, L. E., Scherr, P. A., McCann, J. J., Beckett, L. A., & Evans, D. A. (2001). Is the risk of developing Alzheimer's disease greater for women than for men? *American Journal of Epidemiology, 153*(2), 132–136.

Huang, C., Tian, Y., Zhou, Z., Ling, C. X., & Huang, T. (2006). *Keyphrase extraction using semantic networks structure analysis* (pp. 275–284). Presented at the Sixth IEEE international conference on data mining, ICDM'06.

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., & Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America, 101*(42), 15148–15153.

Kukull, W. A., Higdon, R., Bowen, J. D., McCormick, W. C., Teri, L., Schellenberg, G. D., et al. (2002). Dementia and Alzheimer disease incidence: A prospective cohort study. *Archives of Neurology, 59*(11), 1737–1746.

Lambiotte, R., Delvenne, J. C., & Barahona, M. (2009). Laplacian dynamics and multiscale modular structure in networks. *ArXiv preprint arXiv*: 0812.1770.

Li, J., Zhu, X., & Chen, J. Y. (2009). Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. *PLoS Computational Biology, 5*(7), e1000450. doi:10.1371/journal.pcbi.1000450.

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The unified medical language system. *Methods of Information in Medicine, 32*(4), 281–291.

Miech, R., Breitner, J., Zandi, P., Khachaturian, A., Anthony, J., & Mayer, L. (2002). Incidence of AD may decline in the early 90 s for men, later for women The Cache County study. *Neurology, 58*(2), 209–218.

Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. Presented at the *Proceedings of the 24th conference on uncertainty in artificial intelligence* (pp. 411–418).

Orešič, M., Lötjönen, J., & Soininen, H. (2010). Systems medicine and the integration of bioinformatic tools for the diagnosis of Alzheimer's disease. *Genome Medicine, 2*(11), 83.

Ravetti, M. G., Rosso, O. A., Berretta, R., & Moscato, P. (2010). Uncovering molecular biomarkers that correlate cognitive decline with the changes of hippocampus' gene expression profiles in Alzheimer's disease. *PLoS One, 5*(4), e10153. doi:10.1371/journal.pone.0010153.

Rocca, W. A., Cha, R. H., Waring, S. C., & Kokmen, E. (1998). Incidence of dementia and Alzheimer's disease: A reanalysis of data from Rochester, Minnesota, 1975–1984. *American Journal of Epidemiology, 148*(1), 51–62.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Seshadri, S., Wolf, P., Beiser, A., Au, R., McNulty, K., White, R., et al. (1997). Lifetime risk of dementia and Alzheimer's disease: The impact of mortality on risk estimates in the Framingham Study. *Neurology, 49*(6), 1498–1504.

Shehata, S., Karray, F., & Kamel, M. (2007). A concept-based model for enhancing text categorization. Presented at the *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 629–637). New York: ACM.

Smalheiser, N. R., & Swanson, D. R. (1996). Linking estrogen to Alzheimer's disease: An informatics approach. *Neurology, 47*(3), 809–810.

Smalheiser, N. R., & Swanson, D. R. (1998). Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine, 57*(3), 149–153.

Song, M., Kim, S., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. *Journal of the Association for Information Science and Technology, 65*(2), 352–371. doi:10.1002/asi.22970.

Sorensen, A. A. (2009). Alzheimer's disease research: scientific productivity and impact of the top 100 investigators in the field. *Journal of Alzheimer's Disease, 16*(3), 451–465.

Sorensen, A. A., Seary, A., & Riopelle, K. (2010). Alzheimer's disease research: A COIN study using co-authorship network analytics. *Procedia-Social and Behavioral Sciences, 2*(4), 6582–6586. doi:10.1016/j.sbspro.2010.04.068.

Thota, H., Rao, A. A., Reddi, K. K., Akula, S., Changalasetty, S. B., & Srinubabu, G. (2007). Alzheimer's disease care and management: Role of information technology. *Bioinformation, 2*(3), 91–95.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*(1), 141–188.

Wan, X., Yang, J., & Xiao, J. (2007). *Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction* (Vol. 45(1), p 552). Presented at the Annual Meeting-Association for Computational Linguistics.

Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge: Cambridge University.