A Fuzzy Approach for Measuring Development of Topics in Patents Using Latent Dirichlet Allocation

Hongshu Chen [a,b], Guangquan Zhang [a], Jie Lu [a]
Decision Systems & e-Service Intelligence Lab [a]
Centre for Quantum Computation & Intelligent Systems
Faculty of Engineering and Information Technology
University of Technology Sydney
Sydney, Australia
hongshu.chen@student.uts.edu.au, {guangquan.zhang, jie.lu}@uts.edu.au

Donghua Zhu [b]
School of Management and Economics [b]
Beijing Institute of Technology
Beijing, China
zhudh111@bit.edu.cn

Abstract*—Technology progress brings the very rapid growth of patent publications, which increases the difficulty of domain experts to measure the development of various topics, handle linguistic terms used in evaluation and understand massive technological content. To overcome the limitations of keyword-ranking type of text mining result in existing research, and at the same time deal with the vagueness of linguistic terms to assist thematic evaluation, this research proposes a fuzzy set-based topic development measurement (FTDM) approach to estimate and evaluate the topics hidden in a large volume of patent claims using Latent Dirichlet Allocation. In this study, latent semantic topics are first discovered from patent corpus and measured by a temporal-weight matrix to reveal the importance of all topics in different years. For each topic, we then calculate a temporal-weight coefficient based on the matrix, which is associated with a set of linguistic terms to describe its development state over time. After choosing a suitable linguistic term set, fuzzy membership functions are created for each term. The temporal-weight coefficients are then transformed to membership vectors related to the linguistic terms, which can be used to measure the development states of all topics directly and effectively. A case study using solar cell related patents is given to show the effectiveness of the proposed FTDM approach and its applicability for estimating hidden topics and measuring their corresponding development states efficiently.

Keywords*—fuzzy set; Latent Dirichlet Allocation; topic modelling; patent claims

I. INTRODUCTION

Nowadays, affected by rapid technology progress, patents are applied and issued more than ever before, which produce massive amounts of textual data containing valuable technological overview and details [1]. However, manually conducting content analysis and evaluation on a mass of technical terms is very time consuming and laborious. Thus automatic approaches for revealing and analysing topics hidden in large numbers of patent documents are in great demand [2].

To achieve the aim of understanding valuable thematic knowledge from massive textual data, there are two main phases need to be considered. The first one is automatically learning topics, and the second one is assisting further thematic evaluation using the estimated topics. Much effort has already been devoted to discover latent knowledge from the textual data of patent documents using text mining, for the first phase. For example, Watts and Porter [3] suggested an approach to investigate terminological trends by tracking the historical change of keywords; Yoon and Park [4] proposed a keyword-based morphology research to identify the detailed configurations of promising technology; Cascini and Russo [5] introduced a computer-aided approach based on text mining for accomplishing TRIZ less time-consuming analysis. Generally speaking, the outcomes of traditional text mining techniques applied to assist topic extraction are mostly single keywords with ranking. These words alone, however, are usually too general or misleading to indicate a concept, especially when there are polysemous words actually describing different topics [2]. Moreover, for the second phase of assisting further thematic evaluation, simple term frequency may not be sufficient to support the measurement of development states that various topics have. Especially, the evaluation result in a real case is often expected to be a group of linguistic terms, such as 'growing', 'stable', 'have potential' and so forth, other than numerical values. Under such circumstance, approaches that are capable of discovering multiple words topics automatically and dealing with the vagueness of linguistic terms are needed.

In a real situation, as mentioned, the judgement on certain states, relations or tendency are often expressed by linguistic terms. To deal with the vagueness nature of these terms and manipulate imprecise values in real life, fuzzy sets were introduced by Zadeh [6] as a classical notation of 'set' extension. Since fuzzy sets can effectively handle linguistic terms in measurement and deal with the uncertainty, in this research, we propose a fuzzy set-based topic development measurement (FTDM) approach to estimate and evaluate the topics hidden in a large volume of patent claims using Latent Dirichlet Allocation (LDA). Latent semantic topics are generated with LDA, which utilizes a probability distribution over words, instead of a single term, to define a concept. Thus polysemy is allowed and the semantic meaning of topics is better delivered. Then a temporal-weight matrix is defined to measure the importance of all topics in different years. Based on this matrix, we then calculate a temporal-weight coefficient for each topic, which can be seen as a numerical membership value that associated with a linguistic term, to describe its development states over time. To demonstrate the effectiveness of the proposed FTDM approach, a case study using solar cell

related patents is presented. The result provides the development states evaluation for 30 estimated semantic topics in this area, which shows the applicability of our proposed approach in efficiently estimating hidden topics and measuring their corresponding development states.

This paper is organized as follows. Section II reviews the background of patent claims, the definition of fuzzy sets, and Latent Dirichlet Allocation. Section III presents the fuzzy approach of dealing with development states measurement of technological topics. In Section IV, a case study using United States patents is provided. Section V concludes this research with a discussion and gives future study.

II. PRELIMINARIES

A. Patent Claims

As an important part of unstructured segments of a patent document, claims embody all the most essential technological terms and topics to define the protection of an invention [7, 8]. They reflect to the core inventive idea of a patent, at the same time they provide a direct technological scope in which patent examiners classify the patent to different technological classes [9]. To satisfy specific legal requirements, patent claims shall be concise and clear, and are always described in precise statement [10, 11], which make them the best resource for technological topic mining and content analysis. A patent claim usually consists of three parts: a Preamble that serves as an introductory part to recite the primary purpose, function or properties; a transition phrase, such as comprising, having including, consisting of, etc.; a "body" contains the elements or steps that together describe the invention [7, 12, 13]. Among patent databases from different countries, the United States Patent and Trademark Office (USPTO) database is mostly used for standardization reasons.

B. Fuzzy Sets

Fuzzy sets are characterized by membership functions that assign each object to a grade of membership ranging from 0 to1 [14]. The mathematical definition and notations of a fuzzy set is reviewed from Zhang and Lu [15, 16] as follows:

Definition 1. Let X be a universe discourse. A fuzzy set $\tilde{A}$ in X is characterized by its membership function $\mu_{\tilde{A}}(x)$.

$$x \vdash \mu_{\tilde{A}}(x) \in [0,1] \tag{1}$$

where the membership function $\mu_{\tilde{A}}(x)$ associates with each element x in X a real number in the interval of $[0,1]$. This real number is interpreted as the grade of x belongs to $\tilde{A}$. That is, the closer the value of $\mu_{\tilde{A}}(x)$ is to 1, the more it belongs to the fuzzy set $\tilde{A}$. A fuzzy set can be presented as a set of ordered pairs of elements x and its corresponding grade $\mu_{\tilde{A}}(x)$, which is noted by,

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) | x \in X\}. \tag{2}$$

Definition 2. A fuzzy set $\tilde{A}$ in a universe of discourse X is convex if and only if for any $x_1, x_2 \in X$,

$$\mu_{\tilde{A}}(\lambda x_1 + (1-\lambda)x_2) \geqq \min(\mu_{\tilde{A}}(x_1), \mu_{\tilde{A}}(x_2)), \tag{3}$$

where $\lambda \in [0,1]$.

Definition 3. A fuzzy set $\tilde{A}$ in a universe of discourse X is called a normal fuzzy set implying that there exists $x_0 \in X$ such that $\mu_{\tilde{A}}(x_0) = 1$.

Definition 4. A fuzzy number $\tilde{a}$ is a fuzzy subset on the space of real number R that is both convex and normal. A triangular fuzzy number $\tilde{a}$ can be defined by a triplet (a_0^L, a, a_0^R) and the membership function $\mu_{\tilde{a}}(x)$ is defined as:

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{(x-a_0^L)}{(a-a_0^L)}, a_0^L \leq x \leq a, \\ \frac{(a_0^R-x)}{(a_0^R-a)}, a \leq x \leq a_0^R, \\ 0, \quad others. \end{cases} \tag{4}$$

Definition 5. A linguistic variable is a variable whose values are linguistic terms, such as 'good', 'stable', 'young' and 'old'.

In this research, three sets of linguistic terms are utilized to describe the developing states of estimated topics case by case. For technological area appears strong growing potential, we use terms I={Steady (IS), Gradual Increasing (GI), Rapid Growing (RG)}; for technologies that have been comparatively mature, we choose a term set D={Rapid Declining (RD), Gradual Declining (GD), Steady (DS)}; for technologies show a wave-type of development, we use term set W={Declining (WD), Steady (WS), Growing (WG)}.

C. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [17] is a probabilistic model that uses unsupervised learning to estimate the properties of multinomial observations. It provides an estimation of the latent semantic topics hidden in massive documents. In addition, it also estimates the probabilities of how various documents belong to different topics [18]. In practice, LDA has been utilized as a very efficient tool to assist topic discovery. For instance, Griffiths and Steyvers [19] applied LDA-based topic modelling to discover the hot topics covered by papers in Proceedings of the National Academy of Sciences of the United States of America; Yang and his colleagues [20] proposed a Topic Expertise Model (TEM) for Community Question Answering with Stack Overflow data based on LDA to jointly model topics and expertise; Ding [21] introduced topic-dependent ranks based on the combination of a topic model and a weighted PageRank algorithm. Fig. 1 presents the graphical model of LDA, showing three rectangular plates, where: D denotes the overall documents in a corpus; K indicates the topic numbers for D; and N_d stands for the term number of d^{th} document in the collection D.

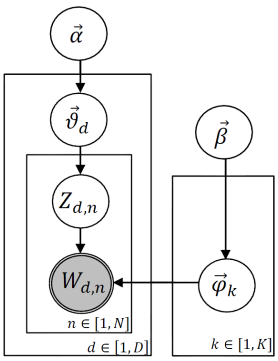

Fig. 1. The graphical model of Latent Dirichlet Allocation

In the generative process of LDA showing in the figure above, each node stands for a random variable and all the three plates indicate replication. On the left of the figure, $\vec{\vartheta}_d$ stands for the topic proportions for the d^{th} document. For document d, the topic assignments are Z_d, where $Z_{d,n}$ indicates the topic assignment of the n^{th} word in the d^{th} document. On the right of the figure, the topics themselves are illustrated by $\vec{\varphi}_{1:K}$, where each $\vec{\varphi}_k$ is a distribution over vocabularies. The shaded circles are observable nodes, where $W_{d,n}$ stands for the n^{th} word in document d. All of the unshaded circles indicate hidden nodes. Finally, α and β are two hyperparameters that determine the amount of smoothing applied to the topic distributions for each document and the word distributions for each topic [17, 22, 23]. In summary, the generative process of LDA can be denoted by the joint distribution of the random variables as follows,

$$p(\vec{w}_d, \vec{z}_d, \vec{\vartheta}_d, \phi|\vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_d} p(w_{d,n}|\vec{\varphi}_{z_{d,n}}) p(z_{d,n}|\vec{\vartheta}_d) p(\vec{\vartheta}_d|\vec{\alpha}) p(\phi|\vec{\beta}). \quad (5)$$

The required parameters of LDA need to be estimated by an iterative approach. Among existing approaches, Gibbs sampling, which is one of the most commonly used methods, is an approximate inference algorithm based on the Markov Chain Monte Carlo (MCMC) and widely used to estimate the assignment of words to topics by observed data [19, 24].

III. METHODOLOGY

This section explains the details of our proposed fuzzy set-based topic development measurement (FTDM) approach.

A. Framework

The framework of the FTDM approach is shown in Fig. 2. After a target technological area is determined, all patents belong to the scope are crawled to a database waiting for further analysis. Then the titles and claims of patent documents, and their corresponding patent ID and Issue dates, are extracted separately. The claims and title of each patent constitute one textual document in our corpus, while the patent ID and Issue Date of all patents compose a single document. On one hand, the textual data are passed to several segmentation and cleaning modules to remove all the punctuations, meaningless symbols, stop-words, general words using in claims and high frequency academic words. Subsequently, Latent Dirichlet Allocation is utilized to generate latent topics from the prepared corpus. On the other hand, the result of topic modelling and the patent Issue Date information are gathered to serve the topic weight estimation. Because patents are published following time order, a list of patent assigned by Issue Date is actually following a strict time line. A temporal-weight matrix is then created to illustrate how the weights of all topics change over time. Then a group of linguistic terms are decided for depicting different states of topic development. According to the outcome of the topic weight estimation module, we can create a fuzzy membership function case by case. For each topic, a temporal-weight coefficient is calculated, which is associated with a set of linguistic terms to describe its development state over time. After choosing a suitable linguistic term set, fuzzy membership functions are created for each term. Finally, the temporal-weight coefficients can be transformed to membership vectors related to the linguistic terms in the module of fuzzy set-based topic development measurement.

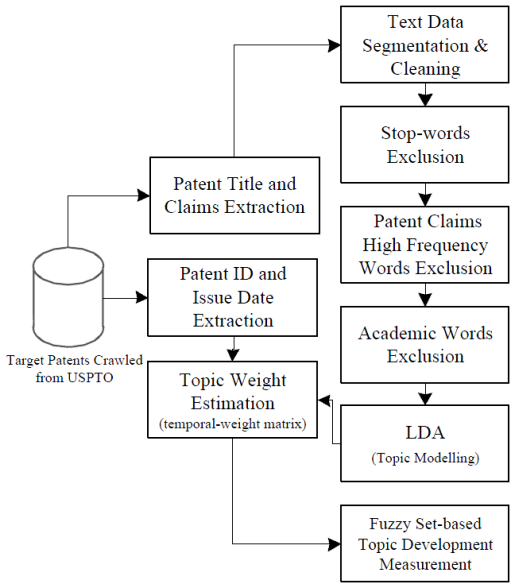

Fig. 2. The framework of FTDM approach

B. Patent Claim Volumes Cleaning

Patent claims are a special kind of textual data that contain plenty of technical terms, specific words serving as transition phrases and numerous academic words that describe invention outcomes. Among all the terms that one claim may contain, only technical terms provide most meaningful information reflecting technological topics. Therefore, for patent collections of each year, as shown in Fig. 2, before modelling topics with LDA, we utilize three modules to remove general words from the corpus of patents as follows:

- Stop-words such as the, that, these;
- High frequency words in patent claims such as claimed, comprising, invention;
- General academic words such as research, approach, data.

The stop words list we applied is from an information retrieval Resources link from Stanford University [25]; the patent claim commonly used phrases are summarized from a Transitional Phrase page on Wikipedia [26]; the general academic words list is provided by the University of Nottingham, we select the top 100 most frequent academic words and remove them from our final corpus [27].

C. Topic Modelling

As a probabilistic model for unsupervised learning, LDA generates a probability distribution over words, instead of a single term, to define a concept, delivering the semantic meaning of the topic. Thus polysemy is allowed, since a same word can serve different topic with different probabilities. After removing all commonly used words from the corpus, we utilize LDA to generate K topics in D documents in our prepared corpus.

Generally speaking, we know nothing about the word distributions composing the topics and the topic distributions

composing the documents, thus assumptions need to be first drawn to determine the parameters k, α, β of LDA. Hyper-parameters α, β of the Dirichlet distribution in LDA have a smoothing effect on multinomial parameters; that is, the lower the values of α and β are, the more decisive topic associations there will be [23]. This research sets $\alpha = 0.5$ and $\beta = 0.1$, which are commonly used in LDA applications. For the setting of K, during the implementation, K needs to be decided case by case to balance user requirement and time consumption, since higher K will reduce the topical granularity but increase the processing time. Different parameter settings may improve modelling performance, yet optimizing these parameters is beyond the scope of this paper. We then apply Gibbs sampling to infer the needed distributions in LDA.

D. Topic Weight Estimation

As an important part of LDA outcomes, the topic distribution matrix provides the estimated result that how all the topics distribute over the document collection. As mentioned, we get D documents express K topics totally, which means we have a topic distribution matrix with D rows and K columns. Each row of the topic distribution indicates how different topics distribute over a document in the corpus. Thus the summation of every row equals 1. The sum values of each column, however, are different. The larger the sum of a column, the more important the corresponding topic is. Since the patents are issued following a time line, if we add up a group of elements in a column that associates with patents published in a same year, the summation can be used to present the weight of the topic in that year. Fig. 3 shows the example of topic distribution matrix.

	Topic 1	Topic 2	Topic 3	•••	Topic K	
Document 1	0.0132	0.0161	0.0161	•••	0.0015	Year 1
Document 2	0.0002	0.0002	0.0008	•••	0.0005	
⋮	⋮	⋮	⋮	•••	⋮	
Document 44	0.0001	0.0001	0.0001	•••	0.0029	
Document 45	0.0001	0.0245	0.0003	•••	0.0003	Year 2
⋮	⋮	⋮	⋮	•••	⋮	
Document 98	0.0001	0.0001	0.0001	•••	0.0128	⋮
⋮	⋮	⋮	⋮	•••	⋮	
Document D	0.0520	0.0020	0.0012	•••	0.0004	Year T

Fig. 3. The example of topic distribution matrix

We got D documents that published during T years. For each topic, every year we have a coefficient equals to the sums of corresponding column elements. Thus, as shown in Fig. 4, we can get a temporal-weight matrix with T rows and K columns to reveal the importance of all topics in different years.

	Topic 1	Topic 2	Topic 3	•••	Topic K
Year 1 weight	2.0172	1.6190	0.4411	•••	0.5875
Year 2 weight	1.3428	2.4317	1.6732	•••	0.8146
Year 3 weight	1.1954	2.7350	0.4394	•••	0.8069
⋮	⋮	⋮	⋮	•••	⋮
Year T weight	11.6390	16.5351	4.6622	•••	7.5101

Fig. 4. The example of temporal-weight matrix

E. Fuzzy Set-based Topic Development Measurement

After a temporal-weight matrix is estimated, we calculate the weight changing rate of each topic using the method of least-squares, which fit each column in the matrix to a straight line. We then define the first coefficient in one degree polynomial is the temporal-weight coefficient of the corresponding topic. For all the generated topics, we get a temporal-weight coefficient vector TW, where $TW = (tw_1, tw_2, tw_3, \dots, tw_K)$ and tw_K stands for the coefficient of the K^{th} topic.

The vector TW is actually an attribute that associates with a set of linguistic terms to describe the development states of all estimated topics. We know that, in existing research, the type of membership function that is suitable depends on the application context [28, 29]. In this research, fuzzy membership functions can be inferred from the analysis of TW, or they may be determined by domain experts. Specifically, a domain value s is determined bases on the domain of TW, where $s \geq |TW_{max} - TW_{min}|$. In addition, a value b is set in each function as a buffer. Because the TW will change for topics in different technological area, the membership functions for the linguistic terms associated with the attribute will be changed accordingly. In this research, we use fuzzy numbers to present different terms.

The shapes of the membership functions of three sets of linguistic terms are illustrated in Fig. 5, in which the sub-figure a shows the membership function of term set I; the sub-figure b illustrates the membership function of term set D; the sub-figure c provides the membership function of term set W.

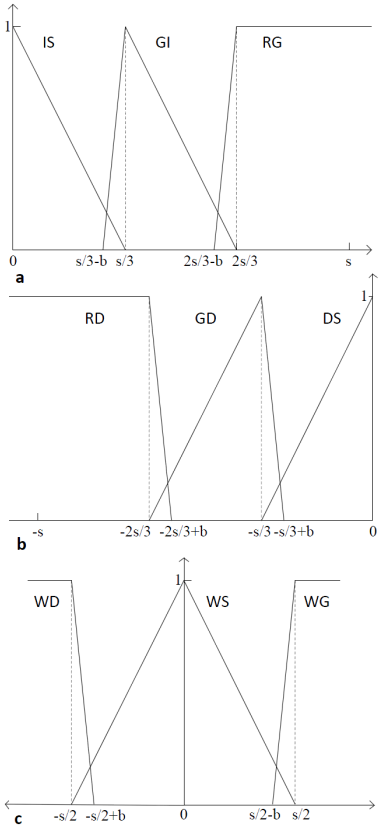

Fig. 5. Linguistic terms and their membership functions

A mapping between TW and linguistic terms that describe topics development tendency can be built then. Each mapping denotes what development state a topic is on. The fuzzy numbers related to the linguistic terms IS, GI, DS, GD and WS are shown in Table I.

TABLE I. LINGUISTIC TERMS AND FUZZY NUMBERS

Linguistic Terms	Fuzzy Numbers
IS	(0, 0, s/3)
GI	(s/3-b, s/3, 2s/3)
DS	(-s/3, 0, 0)
GD	(-2s/3, -s/3, -s/3+b)
WS	(-s/2, 0, s/2)

For linguistic terms RG, RD, WD and WG, however, it's hard to use fuzzy numbers to build the mapping. In this research, fuzzy membership functions $\mu_{RG}(x), \mu_{RD}(x), \mu_{WD}(x)$ and $\mu_{WG}(x)$ are defined respectively for these linguistic terms as follows:

$$\mu_{RG}(x) = \begin{cases} 0, & x < 2s/3 - b \\ \frac{x}{b} + 1 - \frac{2s}{3b}, & 2s/3 - b \le x \le 2s/3 \\ 1, & x > 2s/3 \end{cases} \quad (6)$$

$$\mu_{RD}(x) = \begin{cases} 1, & x < -2s/3 \\ -\frac{x}{b} + 1 - \frac{2s}{3b}, & -2s/3 \le x \le -2s/3 + b \\ 0, & x > -2s/3 + b \end{cases} \quad (7)$$

$$\mu_{WD}(x) = \begin{cases} 1, & x < -s/2 \\ -\frac{x}{b} + 1 - \frac{s}{2b}, & -s/2 \le x \le -s/2 + b \\ 0, & x > -s/2 + b \end{cases} \quad (8)$$

$$\mu_{WG}(x) = \begin{cases} 0, & x < s/2 - b \\ \frac{x}{b} + 1 - \frac{s}{2b}, & s/2 - b \le x \le s/2 \\ 1, & x > s/2 \end{cases} \quad (9)$$

After observing vector *TW*, we can determine what type of technological development the target area has, and select the most suitable term set from I, D and W, for further topic development states analysis.

IV. CASE STUDY AND DISCUSSION

In order to demonstrate the effectiveness of the proposed FTDM approach, we choose solar cell area as a case study, to discover the development states of various detailed topics in it. We collect all the patents related to solar cell (ABST/"solar cell") and published during years 1985 to 2014 in USPTO (http://www.uspto.gov/). Totally, there are 3277 target patents covering 3271 utility patents, 5 reissue patents and 1 statutory invention registration in solar cell area. Their patent ID, titles, issue date and claims are crawled from USPTO and placed in a patent database for further processing. Patents ID and the issue time of theirs are put into one single document, while the claims and title for each patent constitute one document in our corpus, which totals 3277 documents in all. While volume cleaning, besides all the general and meaningless words, we also exclude the word 'solar' and 'cell', which are the highest frequency words in this area.

Before topic modelling, as mentioned, a number of parameters need to be set first, including the number of topics, α, β of Dirichlet distribution and the number of iterations for Gibbs sampling. In this case study, we applied $K = 30$ with model hyper-parameters $\alpha = 0.5$, $\beta = 0.1$ and 2000 iterations of Gibbs sampling to our target document collection, to balance the topical granularity, convenience of understanding, and the speed of processing. Totally, we got 34607 unique terms in our final corpus for topic modelling.

After topic modelling, we got 30 latent semantic topics, in which each of them is presented by the top 20 ranked words and their corresponding probabilities. We then generated the temporal-weight matrix and temporal-weight coefficients, TW, for all topics based on the topic distribution matrix. Table II shows the temporal-weight coefficients of all 30 topics sorted from the largest to the smallest. The larger the TW coefficient is for a topic, the more rapid it is developing.

TABLE II. THE TEMPORAL-WEIGHT COEFFICIENTS OF TOPIC 1 TO TOPIC 30

Topic	TW coefficient	Topic	TW coefficient
Topic 14	1.3451	Topic 4	0.2741
Topic 23	0.8643	Topic 20	0.2121
Topic 7	0.6482	Topic 13	0.1911
Topic 18	0.6354	Topic 1	0.1816
Topic 26	0.5562	Topic 17	0.1761
Topic 6	0.4509	Topic 29	0.1417
Topic 24	0.4416	Topic 30	0.1339
Topic 22	0.4009	Topic 19	0.1214
Topic 28	0.3684	Topic 15	0.1133
Topic 8	0.3430	Topic 9	0.1094
Topic 21	0.3343	Topic 5	0.1032
Topic 12	0.2990	Topic 3	0.0833
Topic 25	0.2965	Topic 11	0.0803
Topic 2	0.2863	Topic 27	0.0650
Topic 10	0.2776	Topic 16	0.0646

After observing the scope of the domain of TW, linguistic term set I is selected as the suitable one, since all the topics showed growing potential. For fuzzy membership functions creation, we set $s = 1.5$ and $b = 0.2$ in this case study. Fig.6 illustrates the final membership functions for term set I.

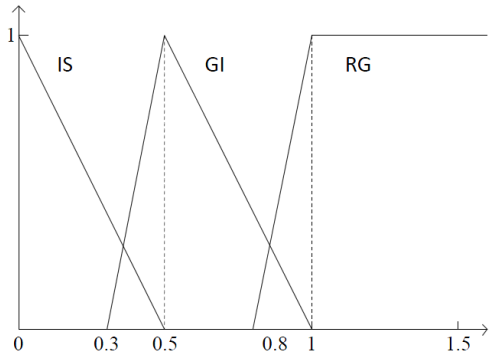

Fig. 6. Membership functions of linguistic terms in term set I

To measure the development state of a semantic topic, we then map each topic coefficient to a linguistic term, by calculating its fuzzy membership degree vector (FMDV). In table III, we listed 15 topics of the final result of using FTDM on solar cell related patents, which have the highest temporal-weight values. For each topic, its fuzzy development measurement and 5 top ranked words are illustrated. For each word, there is a probability value indicating how possible this

word belongs to its current topic. We can see from the form that topic No. 14 that relates to 'silicon substrate', and topic No.23 that concerns 'oxide polymer precursor' are the most rapid growing topic on the whole. Its development state can be measured as 'Rapid Growing'. Topics No. 7, No. 18, No. 26, No. 6, No. 24, No. 22 and No. 28 are measured as 'Gradual Increasing'. The rest of topics, including the topics does not show in the form, are all 'Steady', which means their growing potential are not as strong as the topics mentioned above.

TABLE III. DEVELOPMENT STATES MEASUREMENT RESULT

Topic	Words	RROB	FMDV	Linguistic Terms
Topic 14	Silicon Substrate Surface Dopant metal	0.0564 0.0430 0.0320 0.0227 0.0154	(0,0,1)	RG (Rapid Growing)
Topic 23	metal oxide solution precursor polymer	0.0207 0.0182 0.0162 0.0141 0.0128	(0,0.27, 0.32)	RG (Rapid Growing)
Topic 7	conversion photoelectric glass organic paste	0.0511 0.0358 0.0316 0.0236 0.0186	(0, 0.7, 0)	GI (Gradual Increasing)
Topic 18	electrode plurality surface rear substrate	0.0846 0.0360 0.0274 0.0245 0.0139	(0, 0.73, 0)	GI (Gradual Increasing)
Topic 26	substituted formula unsubstituted compound alkyl	0.0230 0.0191 0.0180 0.0156 0.0127	(0, 0.89, 0)	GI (Gradual Increasing)
Topic 6	film transparent thin substrate conductive	0.1102 0.0517 0.0498 0.0432 0.0377	(0.1,0.75,0)	GI (Gradual Increasing)
Topic 24	power voltage circuit control signal	0.0705 0.0497 0.0220 0.0205 0.0174	(0.12,0.71,0)	GI (Gradual Increasing)
Topic 22	acid encapsulant copolymer composition weight	0.0220 0.0199 0.0137 0.0119 0.0118	(0.2,0.5,0)	GI (Gradual Increasing)
Topic 28	housing upper mounting body plate	0.0173 0.0136 0.0122 0.0116 0.0108	(0.26,0.34,0)	GI (Gradual Increasing)
Topic 8	conductive contact structure electrical surface	0.1054 0.0722 0.0431 0.0278 0.0267	(0.31,0.22,0)	IS (Steady)
Topic 21	surface radiation reflective optical concentrator	0.0586 0.0308 0.0308 0.0209 0.0176	(0.33,0.17,0)	IS (Steady)
Topic 12	system wireless source communicatio n	0.0246 0.0190 0.0180 0.0169 0.0165	(0.4, 0, 0)	IS (Steady)
Topic 25	surface upper semiconductor barrier silicon	0.0240 0.0193 0.0155 0.0154 0.0152	(0.41, 0, 0)	IS (Steady)
Topic 2	plurality elongated absorber internal transparent	0.0320 0.0252 0.0214 0.0159 0.0149	(0.42, 0, 0)	IS (Steady)
Topic 10	panel array plurality support located	0.0357 0.0253 0.0231 0.0153 0.0130	(0.44, 0, 0)	IS (Steady)

It is known that, as a green energy, the solar cell has experienced very vigorous growth during the past decade. The consumption of solar cell technologies continues to rise from the market's perspective [30]. The topic development measurement result on the whole shows the same tendency. More specifically, from the result we can know that the importance of topic "silicon substrate" and "oxide polymer precursor" are growing most rapidly, followed by topics on "photoelectric conversion", "plurality electrode substrate", "alkyl compound", "conductive thin substrate", "voltage and circuit", "acid encapsulant" and "housing (shell) mounting". These topics are more active than other topics in solar cell area. The developing potential of theirs, are stronger. The result of FTDM provides domain experts a direct view and a foresight in topics themselves and their corresponding development in a target area. In summary, FTDM can be used to obtain the main topics and their development states automatically from a large volume of documents, which makes it possible to set domain experts and analysts free from the heavy work of understanding and evaluating massive technological content.

V. CONCLUSION AND FUTURE STUDY

In this research, a FTDM approach is developed based on LDA, to overcome the limitations that keyword-ranking type of text mining result may bring, and at the same time deal with the vagueness of linguistic terms to assist further thematic evaluation. Semantic topics are generated with LDA, which utilizes probability distributions over words to define a concept, instead of using single terms, thus polysemy is allowed. Then a temporal-weight matrix is defined to measure the importance of all topics in different years. Fuzzy membership functions are built for three sets of linguistic terms in this research, to deal with technological areas that are on different phases of their life cycles. Then based on the temporal-weight matrix, for each topic, a temporal-weight coefficient is calculated, which is associated with a set of linguistic terms to describe its development state over time. After choosing a suitable linguistic term set, the temporal-weight coefficients are transformed to membership vectors related to the linguistic terms, which can be used to measure the development states of all topics directly and effectively. A case study using solar cell related US patents is presented in this research to demonstrate the effectiveness of the proposed FTDM approach. The result provides the development states evaluation for 30 estimated semantic topics in this area, which shows the applicability of our proposed approach in efficiently estimating hidden topics and measuring their corresponding development states.

As patents and other technical indicators are still generating and accumulating in an increasing rate, approaches for automatically identifying and analysing latent topics will continue to be emphasized. The FEDM approach can also be used in understanding and evaluating scientific literatures. In future work, we will continue to focus on estimating topic development that associate with more meaningful temporal segmentation, like trend turning intervals [31], to evaluate and analyse the development of trend turning topics.

ACKNOWLEDGEMENT

The work presented in this paper is partly supported by the Australian Research Council (ARC) under Discovery Project DP140101366 and the National High Technology Research and Development Program of China (Grant No. 2014AA015105).

REFERENCES

[1] Wang, W.M. and C.F. Cheung, "A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis". Engineering Applications of Artificial Intelligence, 24(8), 2011, pp. 1510-1520.

[2] Tseng, Y.-H., C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis". Information Processing & Management, 43(5), 2007, pp. 1216-1247.

[3] Porter, A.L., et al., Forecasting and management of technology, New York: Wiley, 1991.

[4] Yoon, B. and Y. Park, "A systematic approach for identifying technology opportunities: Keyword-based morphology analysis". Technological Forecasting and Social Change, 72(2), 2005, pp. 145-160.

[5] Cascini, G. and D. Russo, "Computer-aided analysis of patents and search for TRIZ contradictions". International Journal of Product Development, 4(1),2007,pp. 52-67.

[6] Zadeh, L.A., Fuzzy sets. Information and control, 8(3),1965, pp. 338-353.

[7] Yang, S. and V. Soo, "Extract conceptual graphs from plain texts in patent claims". Engineering Applications of Artificial Intelligence, 25(4), 2012, pp. 874-887.

[8] Tong, X. and J.D. Frame, "Measuring national technological performance with patent claims data". Research Policy, 23(2), 1994, pp. 133-141.

[9] Novelli, E., "An examination of the antecedents and implications of patent scope". Research Policy, 2014, In press.

[10] Xie, Z. and K. Miyazaki, Evaluating the effectiveness of keyword search strategy for patent identification. World Patent Information, 35(1), 2013, pp. 20-30.

[11] WIPO. Patent Cooperation Treaty (PCT) Article 6: Claims. Claims, 2002. Available from: http://www.wipo.int/pct/en/texts/articles/a6.htm.

[12] USPTO. Manual of Patent Examining Procedure:Claim Interpretation. Patent Laws, Regulations, Policies & Procedures, Chapter 2100, Section 2111, 2012,http://www.uspto.gov/web/offices/pac/mpep/s2111.html

[13] Sheldon, J.G., How to Write a Patent Application, Practising Law Institute, 1995.

[14] Zhang, G., Fuzzy number-valued measure theory, Tsinghua University Press, Beijing, 1998.

[15] Lu, J., Zhu, Y., Zeng, X., Koehl, L., Ma, J., and Zhang, G. "A linguistic multi-criteria group decision support system for fabric hand evaluation," Fuzzy Optimization and Decision Making, (8:4), 2009, pp 395-413.

[16] Zhang, G. and J. Lu, "A linguistic intelligent user guide for method selection in multi-objective decision support systems". Information Sciences, 179(14), 2009, pp. 2299-2308.

[17] Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation", the Journal of machine Learning research, vol.3, 2003,pp. 993-1022.

[18] Blei, D.M., "Probabilistic topic models". Communications of the ACM, 55(4), 2012, pp. 77-84.

[19] Griffiths, T.L. and M. Steyvers, "Finding scientific topics". Proceedings of the National academy of Sciences of the United States of America, 101(Suppl 1), 2004, pp. 5228-5235.

[20] Yang, L., et al. "CQArank: jointly model topics and expertise in community question answering". in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, 2013, ACM.

[21] Ding, Y., "Topic-based PageRank on author cocitation networks". Journal of the American Society for Information Science and Technology, 62(3), 2011, pp. 449-466.

[22] Steyvers, M. and T. Griffiths, Probabilistic topic models, in Latent Semantic Analysis: A road to meaning, Laurence Erlbaum,2007.

[23] Heinrich, G., Parameter estimation for text analysis, Fraunhofer IGD: Darmstadt, Germany, 2005.

[24] Noel, G.E. and G.L. Peterson, "Applicability of Latent Dirichlet Allocation to multi-disk search". Digital Investigation, 2014, in press.

[25] David D. Lewis, Y.Y., Tony G. Rose, Fan Li. SMART stopword list. Journal of Machine Learning Research, 2004. Available from: http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop.

[26] Wikipedia. Transitional phrase, 2014. Available from: http://en.wikipedia.org/wiki/Transitional_phrase.

[27] Haywood, S. Academic Vocabulary, 2003. Available from: http://www.nottingham.ac.uk/alzsh3/acvocab/wordlists.htm.

[28] Pedrycz, W. and F. Gomide, An introduction to fuzzy sets: analysis and design, Mit Press, 1998.

[29] Wu, D., G. Zhang, and J. Lu. "A Fuzzy Tree Similarity Measure and Its Application in Telecom Product Recommendation". in 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013. IEEE.

[30] Tseng, F.-M., et al., "Using patent data to analyze trends and the technological strategies of the amorphous silicon thin-film solar cell industry". Technological Forecasting and Social Change, 78(2), 2011, pp. 332-345.

[31] Chen, H., et al., "A patent time series processing component for technology intelligence by trend identification functionality". Neural Computing and Applications, 26(2), 2015, pp. 345-353.