12. If node t_i is not found in G then
13. Initialize an Instance object I as a node of G, Set I.name=t_i
14. Else
15. Set I=The Instance object named t_i in G
16. Increase I. ioccur by 1
17. if(i==0) then
18. Initialize an outlink $R - t_i$ if not found
19. Increase $R - t_i$weight by 1
20. Set $R - t_i$frominstance = R
21. Set $R - t_i$toinstance = I
22. If(i>0 &i<m) then
23. Get PreI=the instance object with name t_{i-1}
24. Initialize an OutLink $t_{i-1} - t_i$ if not found
25. Increase $t_{i-1} - t_i$. iWeight by 1
26. Set $t_{i-1} - t_i$. toInstance = I
27. If (i==m) then
28. Initialize an OutLink t_i-E if not found
29. Increase t_i-E.iWeight by 1
30. Set t_i-E.toInstance = E
31. Set t_i-E.fromInstance = I
32. Set I.hasWPage = PageID
33. Add term t_i into PageID.Keywords
34. End

TermNetWP can be used effectively not only to model the term sequences in connection with Web-pages, but also to present the co-occurrence relations of terms in the term sequences based on the following features: (i) it allows a term node to have multiple in-links and/or out-links so can easily describe the relationships among terms/nodes in the semantic network, i.e. one node might have previous or next nodes; and (ii) it includes the Web-pages whose titles contain the linked terms so that the meaning of Web-pages can be found through these terms by software agents/systems.

Query about pages mapped to a given term: To apply query for Web-pages of a given domain term $t \in T_{auto}$, that can retrieve WPage instances that are mapped to the term instance t via the 'hasWPage' object property. Refer to this query as Pageauto(t). In order to ensure that the degree of relevance of retrieved pages to domain term t is taken into account in the Web-page recommendation process later, the returned pages are sorted in ascending order of connection weights between the Web-pages and domain term t. A connection weight between a Web-page $d \in D$ and domain term t in TermNetWP O is defined as the total number of links from/to domain term t to/from the domain terms of Web-page d. In order to make better Web-page recommendations, need semantic Web usage knowledge which can be obtained by integrating the domain knowledge model (DomainOntoWP) or the semantic network (TermNetWP) with Web usage knowledge that can be discovered from Web log files using a Web usage mining technique

Conceptual Prediction Model (CPM): In order to obtain the semantic Web usage knowledge that is efficient for semantic-enhanced Web-page recommendation, a conceptual prediction model (CPM) is proposed to automatically generate a weighted semantic network of frequently viewed terms with the weight being the probability of the transition between two adjacent terms based on FVTP. The probability of a transition is estimated by the ratio of the number of times the corresponding sequence of states was traversed and the number of times the anchor state occurred. Based on given a CPM having states , $\{S, t_1, \ldots, t_p, E\}$, and $N = |F|$ is the number of term patterns in F, the first-order transition probabilities are estimated according to the following expressions:

$$\rho_{S,x} = \frac{\partial_{S,y}}{\sum_{y=1}^{N} \partial_{S,y}} (1)$$

Which is the first-order transition probability from the starting state S to state tx,

$$\rho_{x,y} = \frac{\partial_{x,y}}{\partial_x} (2)$$

Which is the first-order transition probability from state tx to ty,

$$\rho_{x,E} = \frac{\partial_{x,E}}{\partial_x} (3)$$

Which is the first-order transition probability from state tx to the final state E.

Let $\rho x, y, z$ be the second-order transition probability, that is, the probability of the transition (ty, tz) given that the previous transition that occurred was (tx, ty). The second-order probabilities are estimated as follows:

$$\rho_{X,Y,Z} = \frac{\partial_{x,y,z}}{\partial_{x,y}} \quad (4)$$

Schema of CPM as an ontology schema consists of classes cNode and cOutLink, and relationship properties between them, namely inLink, outLink and linkTo, where cNode and cOutLink defines the current state node and the association from the current state node to a next state node, respectively. The class cNode has two object properties inLink and outLink refering to cNode and cOutLink, respectively. The number of occurrence of each cNode object is represented by Occur, i.e. ∂x. inLink represents an association from a previous state node, e.g. a previous viewed term, to the state node it belongs to. cOutLink represents an association from the state node to one next state node with a transition probability Prob, e.g. $\rho_{x,y}$Given a set of frequently viewed term patterns, namely FVTP, construct TermNavNet by populating the CPM schema with FVTP. The transition probabilities in the *cOutLinks* can be updated based on the first-order or second-order probability formula, depending on the applied CPM's order. As a result, can obtain a 1st or 2nd order TermNavNet by using the 1st or 2nd-order CPM, respectively.Recommendation strategies, that apply the semantic knowledge base of a given website, which includes the domain ontology of Web-pages (DomainOntoWP) or the semantic network of Web-pages (TermNetWP) and the weighted semantic network of frequently viewed terms of Web-pages within the given website(TermNavNet), to make Web-page recommendations. These recommendations are referred to as semantic enhanced Web-page recommendations

Sentiment Variation with background topics and foreground topics: The emerging events or topics are strongly correlated with sentiment variations. Mining such events/topics is not trivial. Topics discussed before the variation period may continue receiving attention for a long time. Formulate this special topic mining problem as follows: given two document sets, a background set B and a foreground set T, want to mine the special topics inside T but outside B. In mining task, the foreground set T contains tweets and the background set B contains tweets appearing before the variation period. Note that this problem setting is general: it has applications beyond sentiment analysis. FB-LDA finds word distributions to reveal possible reasons, which might not be easy for users to understand. Therefore resort to finding representative tweets that reflect foreground topics learnt from FB-LDA. Propose another generative model called Reason Candidate and Background LDA (RCB-LDA) to accomplish this task. RCB-LDA can simultaneously optimize topic learning and tweet-candidate association. RCB-LDA is an extension of FB-LDA. It will accept a set of reason candidates as input and output the associations between tweets and those reason candidates.

Foreground and Background LDA: filter is applied to all topics existing in the background tweets set, known as background topics, from the foreground tweets set. a generative model FB-LDA is proposed to achieve this goal. FB-LDA has two parts of word distributions: φ_f ($K_f \times V$) and φ_b ($K_b \times V$). φ_f is for foreground topics and φ_b is for background topics. K_f and K_b are the number of foreground topics and background topics, respectively. V is the dimension of the vocabulary. For the background tweets set, FB-LDA follows a similar generative process with the standard LDA [24]. Given the chosen topic, each word in a background tweet will be drawn from a word distribution corresponding to one background topic (i.e., one row of the matrix φ_b). However, for the foreground tweet set, each tweet has two topic distributions, a foreground topic distribution θ_t and a background topic distribution μ_t. For each word in a foreground tweet, an association indicator y_i^t, which is drawn from a type decision distribution λ_t, is required to indicate choosing a topic from θ_t or μ_t. If y_i^t,

the topic of the word will be drawn from foreground topics (i.e., from θ_t), as a result the word is drawn from φ_f based on the drawn topic. Otherwise ($y_i^t = 1$), the topic of the word will be drawn from background topics (i.e., from μ_t) and accordingly the word is drawn from φ_b.With the help of background tweets, tweets coming from the foreground set but corresponding to background topics would make a bigger contribution in background topics learning than in foreground topics learning. The large amount of similar background tweets in the background set would pull them to the background topics. Only tweets corresponding to foreground topics (i.e, emerging topics) will be used to build foreground topics. In this way, background topics will be filtered out and foreground topics will be highlighted in a natural way. Given the hyper parameters$\alpha_\theta, \alpha_\lambda, \alpha_\mu, \beta_f, \beta_b$, the joint distribution is:

$$\mathcal{L} = P(y, z_t, z_b, w_t, w_b | \alpha_\theta, \alpha_\lambda, \alpha_\mu, \beta_f, \beta_b)$$
$$= P(y|\alpha_\lambda) P(z_t^{'}|y = 0; \alpha_\theta) P(z_t^{"}, z_b | y = 1; \alpha_\mu$$
$$P(w_t^{'}|y = 0; z_t^{'}; \beta_f) P(w_t^{"}, w_b | y = 1, z_t^{"}, z_b; \beta_b) \quad (5)$$

Reason Candidate and Background LDA: Apart from FB-LDA, RCB-LDA explains a third document set: reason candidates. Reason candidates are in the form of natural language snippets and represent some specific events. In these research automatically finding the most relevant tweets (i.e., representative tweets) for each foreground topic learnt from FB-LDA, using the following measure:

$$\text{Relevance}(t, k_f) = \quad _{i \in t} \phi_f^{k_{f,i}} \quad (6)$$

where $\phi_f^{k_{f,i}}$ is the word distribution for the foreground topic k_f and i is the index of each non-repetitive word in tweet t. Note that we don't normalize this measure with respect to the length of the tweet, since tweets are all very short and generally have similar lengths. For other kinds of texts, normalization shall be applied. After filtering out junk tweets and merging similar ones, consider the remaining relevant tweets as the reason candidates. Generative process of RCBLDA is similar to that of FB-LDA. It generates the reason candidates set and the background tweets set in a similar way as the standard LDA. The main difference lies in the generative process of the foreground tweets set. Each word in the foreground tweets set can select a topic from alternative topic distributions:

1. foreground is taken from the topic distribution of one candidate;
2. Background topic is drawn from its own background distribution μ_t. Specifically, for each word in tweets y_i^t is chosen from the foreground tweets which are similar to that in FB-LDA.

If $y_i^t = 0$ should choose an association candidate c_i^twhich is chsen from a candidate association distribution γ_t. Then draw a foreground topic from θc_i^t for that word. The generative process for $y_i^t = 1$ is as same as that in FB-LDA. Due to the space limit, we reject some parts of the generative process of RCB-LDA which are same as to those in FB-LDA. Here just present the generative process for foreground tweets set in RCB-LDA.

Given the hyper parameters $\alpha_\theta, \alpha_\gamma, \alpha_\lambda, \alpha_\mu, \beta_f, \beta_b$, the joint distribution is

$$\mathcal{L} = P(y, c, z_c, z_t, z_b, w_t, w_b | \alpha_\theta, \alpha_\lambda, \alpha_\mu, \beta_f, \beta_b)$$
$$= P(y|\alpha_\lambda) P(c|\alpha_\gamma) P(z_c, z_t^{'}|y = 0, c; \alpha_\theta) P(z_c, z_t^{"}, z_b | y = 1; \alpha_\mu$$
$$P(w_c, w_t^{'}|y = 0; z_t^{'}; \beta_f) P(w_c, w_t^{"}, w_b | y = 1, z_t^{"}, z_b; \beta_b) \quad (7)$$

Gibbs sampling is used here because it is easy to extend and it has been proved to be quite effective in avoiding local optima. The sampling methods for the two models are similar to each other.

IV. EXPERIMENTATION RESULTS

Twitter dataset is used to analyze public sentiment variations. The dataset is obtained from the Stanford Network Analysis Platform. It is taken from June 11, 2009 to December 31, 2009 and contains around 476 million tweets. It has around 20-30% of all public tweets published on Twitter during that time period. Experiments on a subset of the dataset, which spans from June 13, 2009 to October

31, 2009. In this work, two targets are chosen to test our methods, "Obama" and "Apple". These two targets represent the political sphere and the business field, where the analysis of sentiment variation is very difficult in decision making. It is used to evaluate the precision and recall of the results mined by FB-LDA with respect to the ground truth. The precision and recall of the results found by FB-LDA are computed as follows: (a) rank foreground topics by their word entropies in ascending order. (b) for a foreground topic, five most related tweets are selected.(c) if the most related tweets contain a tweet in the ground truth, or contain a tweet which is very similar to a tweet in the ground truth, believe that the method finds a correct foreground event.

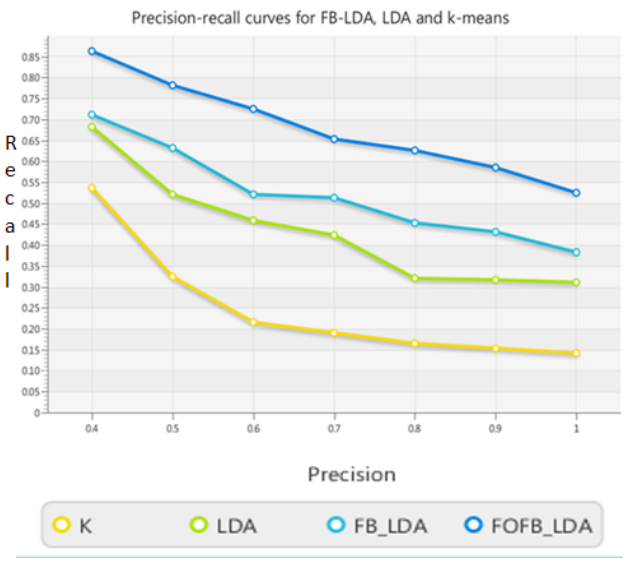

Figure 2: Precision-recall Curves for FB-LDA, LDA and k-means

Figure 2 shows the Precision-Recall curves (average on all 50 variations) for FB-LDA, LDA and k-means. In this experiment FB-LDA is used to produce 20 foreground topics and 20 background topics. For LDA, implement it to produce 20 topics on the foreground set and another 20 topics on the background set. For k-means clustering, two sets are generated respectively, each generating 20 clusters. It is clear that FB-LDA greatly performs well in the two baselines in terms of precision and recall. LDA and k-means depends on threshold and cannot work well due to a fixed threshold is not appropriated for filtering background topics for all cases. In comparison, FB-LDA can work properly without depending on any thresholds.

Table 1: Precision-recall Curves for FB-LDA, LDA and k-means

Precision	K	LDA	FB-LDA	FOFB-LDA
0.4	0.536	0.682	0.712	0.8623
0.5	0.325	0.521	0.6312	0.7814
0.6	0.215	0.4582	0.5213	0.7241
0.7	0.1895	0.4231	0.5131	0.6521
0.8	0.1653	0.3214	0.4523	0.6258
0.9	0.1531	0.3158	0.4321	0.5814
1	0.1425	0.3105	0.3812	0.5238

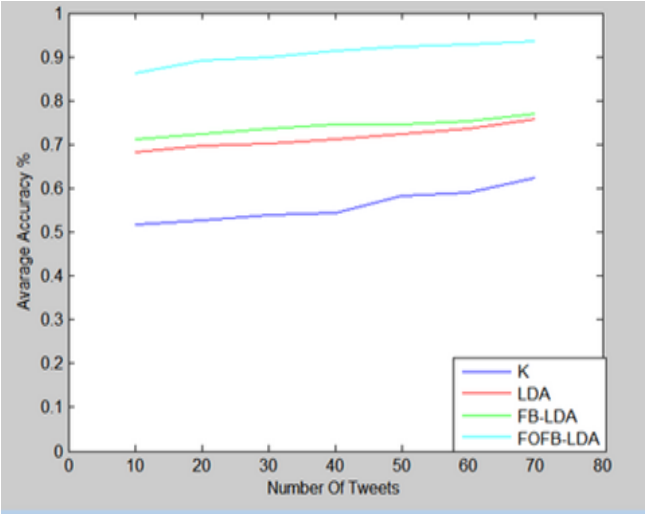

Figure 3: Average Accuracy for FB-LDA, LDA and k-means

Figure 3 shows the comparison of all three models' and average mapping accuracies by varying number of tweets samples .Proposed FOFB-LDA model achieves the best accuracy in a wide range of the parameter variation. Moreover, compared with the two baseline methods, FOFB-LDA method is poor sensitive to the varying threshold. LDA cannot work well for two reasons: (1) the topics learnt by LDA cannot accurately reflect the real foreground events; (2) The optimization of LDA is poor to the association goal directly and the results are illustrated in Table 2.

Table 2: Average Accuracy for FB-LDA, LDA , k-means and Proposed FOFB-LDA Model

Number of tweets	K	LDA	FB-LDA	FOFB-LDA
10	0.516	0.682	0.712	0.8623
20	0.526	0.6981	0.72415	0.8912
30	0.5381	0.7021	0.7351	0.8985
40	0.5428	0.7125	0.7451	0.9125
50	0.5821	0.7245	0.74561	0.9244
60	0.5912	0.7351	0.7531	0.9281
70	0.6231	0.7581	0.7689	0.9356

V. RESULT AND CONCLUSION

In conclusion, research paper presented a new Web-page recommendation method semantic. Two new models have been proposed for representation of domain knowledge of a website. One is an ontology-based model which can be semi-automatically constructed, namely DomainOntoWP, and the other is a semantic network of Web-pages, which can be automatically constructed, namely TermNetWP. A number of Web-page recommendation strategies have been proposed to predict next Web-page requests of users through querying the knowledge bases. To solve the problem, two Latent Dirichlet Allocation (LDA) based models, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA) are proposed. To give a more intuitive representation, the RCB-LDA model can rank candidates expressed in natural language for providing sentence-level reasons. Experimental results showed that proposed models provides possible reasons behind sentiment variations. Moreover, the proposed models are used to discover special topics or aspects in one text collection in comparison with another background text collection. For the future work, a key information extraction algorithm will be developed to compare with the term extraction method, and will perform intense comparisons with the existing semantic Web-page recommendation systems.

REFERENCE

[1] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. International Conference on Weblogs and Social Media (ICWSM), 2008.

[2] Asur, S., and Huberman, B.A. 2010, "Predicting the Future with Social Media", International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, Los Alamitos, CA, USA, 492–499.

[3] Tumasjan, A., Sprenger, T.O., Sandner, P., And Welpe, I.M. 2010, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", paper presented at the 4th International AAAI Conference on Weblogs and Social Media, May 23-26, 2010, Washington, USA.

[4] Bollen, J., Mao, H., Zeng, X.J. 2011, "Twitter mood predicts the stock market", J. Comput. Science, vol. 2, no. 1, 1–8.

[5] Ming Fai Wong, F., Sen, S., and Chiang, M. 2012, "Whay Watching Movie Tweets Won"t Tell the Whole Story?", arXiv preprint, Princeton University, March 21, 2012, available at: http://arxiv.org/abs/1203.4642 (accessed 14 June 2012).

[6] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. International Conference on Weblogs and Social Media (ICWSM). Citeseer, 2007.

[7] B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1-135, 2008. ISSN 1554-0669.

[8] Asiaee T, A., Tepper, M., Banerjee, A., Sapiro, G.: If you are happy and you know it... tweet. international conference on Information and knowledge management. pp. 1602-1606. ACM (2012)

[9] Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T.: Semeval- 2013 task 2: Sentiment analysis in twitter.International Workshop on Semantic Evaluation. Association for Computational Linguistics. (2013)

[10] Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. First workshop on Unsupervised Learning in NLP. Edinburgh, Scotland (2011)

[11] Y. Hu, A. John, F. Wang, and D. D. Seligmann, "Et-lda: Joint topic modeling for aligning events and their twitter feedback," in Proc.26th AAAI Conf. Artif. Intell, Vancouver, BC, Canada, 2012.

[12] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011

[13] B.O'Connor, R. Balasubramanyan, B.R. Routledge , and N. A. Smith, "From tweets to polls: Linking

text sentiment to public opinion time series," AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.

[14] Myriam Munezero, Calkin Suero Montero,Erkki Sutinen, and John Pajunen, "Are They Different? Affect, Feeling, Emotion,Sentiment, and Opinion Detection in Text", IEEE transactions on affective computing, pp. 101-111 VOL. 5, NO. 2, APRIL-JUNE 2014

[15] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[16] Chong Tze Yuang, Rafael E. Banchs, Chng Eng Siong, "An Empirical Evaluation of Stop Word Removal in Statistical Machine Translation" Association for Computational Linguistics, pages 30–37, Avignon, France, April 23 - 27 2012.

[17] G. Antoniou and F. V. Harmelen, A Semantic Web Primer. Cambridge, MA, USA: MIT Press, 2008.

[18] S. Boyce and C. Pahl, "Developing domain ontologies for course content," Educ. Technol. Soc., vol. 10, no. 3, pp. 275–288, 2007.

[19] A. Harth, M. Janik, and S. Staab, "Semantic Web architecture," in Handbook of Semantic Web Technologies, J. Domingue, D. Fensel, and J. A. Hendler, Eds. Berlin, Germany: Springer-Verlag, 2011, pp. 43–75.

[20] S. Grimm, A. Abecker, J. Völker, and R. Studer, "Ontologies and the semantic Web," in Handbook of Semantic Web Technologies, J. Domingue, D. Fensel, and J. A. Hendler, Eds. Berlin, Germany: Springer, 2011, pp. 507–580.

[21] B. Liu, "Information retrieval and Web search," in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, B. Liu, Ed. Berlin, Germany: Springer, 2011, pp. 183–236.

[22] Z. Markov and D. T. Larose, "Information retrieval and web search," in Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage, Z. Markov and D. T. Larose, Eds. New Britain, CT, USA: Wiley, 2007, ch. 1, pp. 3–46.

[23] M. Uschold and M. Gruninger, "Ontologies: Principles, methods and applications," Knowl. Eng. Rev., vol. 11, no. 2, pp. 93–36, Jun. 1996.

[24] Y. Hu, A. John, F. Wang, and D. D. Seligmann, "Et-lda: Joint topic modeling for aligning events and their twitter feedback," in Proc. 26th AAAI Conf. Artif. Intell., Vancouver, BC, Canada, 2012