IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 7 (July. 2018), ||V (I) || PP 06-11

Approaches to personalized news extraction framework

Shine. K. George [1], Jagathy Raj V. P [2]
[1]*Department of Computer Applications, Cochin University of Science and Technology, Kochi, India*
[2] *School of Management Studies, Cochin University of Science and Technology, Kochi, India*
Corresponding Author: Shine. K. George

Abstract: Personalization is inevitable in the field of news domain. Due to the increased use of internet and increase in the number of news channels, the production of news is vast. The public find it difficult to get the latest updates on their interest from the large quantity of news produced. People rely on news stories which consist of related events than the exact news update. The Journalist finds it difficult to extract related news events to create news story from the news archives due to the huge volume of news and insufficient automated tools. We have conducted a study about the existing personalized news extraction approaches for this problem and this paper include its details

Keywords: - Information Extraction, Knowledge Management, Ontology, Personalization, Neural Networks

---

Date of Submission: 21-06-2018 Date of acceptance: 05-07-2018

I. INTRODUCTION

Mass media plays a vital role in communicating with a large audience. Nowadays the media are the fundamental part of our lives. Over the last decade, the amount of information available has grown exponentially. The pace of production is also increasing as more and more news organizations are come into existence and new types of news media hit the market. There are various sources of information and each one seems to produce much more potentially relevant information. As a result of globalization, news from remote parts of the world starts gaining importance, and this can affect some aspects of our lives. Newspapers, television and other media are spreading the news of every event around the world. The average person can read the newspapers and watch a news program on TV. However, it is not possible to know all events in the world. On the other hand, most information from the media resources may not be relevant to the end user. All this is of particular importance to professionals whose work is based on news analysis especially journalist. Computerized systems can help journalist in providing a quick and comprehensive analysis of available sources. The filtering by date and section (politics, economy, etc.) is not enough for today's requirements. There must be automatic work on the body of each news in different types of media to capture the semantics expressed in it. The customization provides a user with a unique view of the information when the information is customized to a user preference. The main benefit of customization is that each user focuses on the information that best suits them. By providing personalized information to a user, they can better handle this information than if they were displayed with the huge general news.

With the development of the Internet, huge information about current news events is continuously generated and disseminated on online news media sites. It is difficult for the public to effectively process such large amounts of information. The automated generation of news summary of certain related news events has recently been intensively studied [1]. The news storyline consists of related news events which are ordered and clustered according to their content and temporal similarities. There are different ways to calculate content and temporal similarities [2,3]. The Bayesian nonparametric models could also be used to solve this problem by describing the storyline generation process using probabilistic graphical models [4]. The existing approaches independently extract news events and link relevant events. The applications of Deep learning techniques are successful in various tasks dealing with natural languages. Therefore, it would be fascinating to merge the benefits of the probabilistic graphical modeling and deep neural networks in news storyline generation. In recent years, some attempts are taken to explore this [5].

This paper is structured as follows. The related works are described in section II. Section III summarizes the review we conducted and finally, in section IV the conclusions and future scope are explained.

II. RELATED WORKS

2.1 Traditional Approach to Personalization

Traditional approaches to customization include the approaches mainly the content-based and user-based [6]. Systems with recommendations are usually rated according to how recommendations are made [7]. A content-based or individual approach requires maintenance of an accurate user profile (eg, the user can provide the system with a sequence of keywords that reflect their original interests, and the profiles can be in the form of weighted and updated keyword vectors based on explicit feedback. However, in the case of areas such as news, a user's interest in an article is not always indicated by the present terms/topics in a document.

In addition to the approaches discussed above, several hybrid approaches have been proposed. These hybrid systems have been motivated by the observation that each of the recommendation technologies developed in the past has certain shortcomings that are difficult to overcome within the bounds of a single recommendation approach. For example, the inability of the approaches like collaborative filtering to recommend new articles can be resolved by linking them to a content-based recommendation approach. Not surprisingly, the most usual form of hybrid recommendation combines content-based and collaborative filtering. There are plenty of recommendation systems based on personalized news are available which follows the traditional approach [8,9,10,11,12,13,14,15,16].

2.2 Ontology-Based approach to Personalization

A traditional approach has one drawback: the difficulty of grasping the semantic knowledge of the area of application, i.e. Concepts, relationships between different concepts, inherent properties associated with concepts, axioms, or other rules, etc. The most crucial thing to learn from these customized news systems is that it is important to have a good understanding of the domain in which the news is available. Generic understanding does not give user-specific results. If there are multiple domains, then a content specific domain model will provide the best result. Domain models that have a good semantic structure will provide better news. The semantic approach to retrieve relevant information can be useful in determining the type or quality of information proposed for a personalized environment [17]. In this context, the standard keyword search has only a very limited effect. For example, it cannot filter the type of information, the level of information, underlying semantics of information or the quality of the information.

An Ontology-Based Information Extraction can be used to draw out the hidden semantics of the news content and can provide a personalized recommendation for the user query. In recent years a lot of ontology-driven news extraction frameworks are developed [18,19,20,21,22,23,24,25,26,27]

One of the major short comings of automatic summarization research is the vague semantic understanding of the source, which is reflected in the poor quality of the resulting abstracts. Using knowledge provided by ontologies to build a semantic representation of the text that is complex can alleviate the problem considerably. In this paper, an extraction method for summarization particularly, ontology-based is presented. It is based on the correlation of text to concepts and the presentation of the document and its sentences as graphics [28]

2.3 language processing tools based approach

Natural language generation (NLG) is known to be the science of "linguistic manipulating of data" [29]. There exist several techniques for NLG including rule-based, statistical and data-driven. Rule-based models rely on hand-written specialized rules for the generation. Rule-based models are also template-based where the models make use of predefined phrases or grammar. The problem with these is that they often require many rules to be crafted and the resulting model will often be restricted to a small specialized domain as well as a small vocabulary. This, in turn, means that the model may perform very well but it lacks flexibility [30]. Statistical models (usually some combination of N-gram models) are based on counts retrieved from the training data. The models allow for explicit modeling of the context and joint probabilities. The issue with these models arises because they are restricted to what they see and their assumption of independence. So these models can't capture any long-term dependencies.

Despite the increased quantity of data as well as the latest evolution of natural language generation techniques, the work in automated journalism is still relatively insufficient. In this article, the challenges associated with the building of a journalistic system of natural language generation are discussed. The architecture for automated journalism that is data-driven specific is proposed and that is independent in domain and language. The paper illustrates its real application in producing articles based on the request of the user in regard to the Finnish municipal elections of 2017 in three languages [31]

As the amount of information consumed daily increases, time to retrieve the information is also important. Staying up to date is an activity that is important to everyone, but also the time you save is important. This paper is primarily focused on implementing techniques that are based on natural language processing and algorithms to aggregate news articles from public sources so that they can be consumed in a short time to keep users informed of global and local events. The proposed solution uses the Text Rank algorithm [32].

Abstractive Summarization generates a shorter advanced version of a text document that is informative by using key Information from the text. This paper proposes an abstract technique and uses natural language processing. The Stanford's approach, as used in this paper, uses NLP instruments, rules of extraction etc. The results prove that the use of Natural Language Processing (NLP) in automatic text compression offer a certain level of abstraction level that is not possible with statistical approaches. The results suggest that based on the parameters Information content, the satisfaction of the reader, length of the summary and grammatical accuracy, proposed approach produces significant results [33].

Today, with the massive demand for sports news, automatic generation systems are being used to the quick and effective generation of massive sports news. In this paper, an automatic generation method based on knowledge rules for dynamically selecting the template from a template set is proposed. It is a data-to-text template-based system that generates summaries from structured input data. The text generated by the system is flexible [34]

2.4 Neural network based approach

Machine generated texts become more and more common over past years. Today there are varieties of methods and techniques are available for producing texts of various types and qualities. As computation power grew with time, the complexity of the models capturing the languages is also increased. Many of the successful models are based on deep learning and neural networks [35,36]

Recently, neural models are used for generating headlines. It was done by using the recurrent neural network which is a learning method that can map documents to the respective headlines. In this study, we include an introduction in detail and a comparison of existing and recent developments in neural headline generation is also explained [37].

The storyline generation aims to extract news events on news articles under a specific news category. Most of the existing approaches are first trained to extract news events that have been published in different periods. Then it will link relevant events into coherent stories. They are domain dependent. To solve this problem, Approaches based on probabilistic graphics models are also jointly used. In this paper, we propose an Extraction framework based on neural networks. The proposed model was assessed on three news corpora [38]

When the recurrent neural networks (RNN) are developed, different natural language generation (NLG) tasks are easily solved in recent years. The NLG fails to study the automated extraction and generation of comments in the news which is a challenging and a new approach. Different from other NLG tasks, this approach needs a contextual meaning. Also, we have to generate comments from varied opinions because for the same news different people across the may have various opinions from their perspective. In this paper, we suggest a Gated Attention Neural Network (GANN) model to in order to generate the comments in the news. We introduce the gated attention technique to self-adaptively and selectively use the context of news so as to tackle the issue of contextual relevance, To ensure the variety of comments, the suggested model uses random sampling and relevance controls to generate comments with different topics [39]

An important role of news sites is to recommend articles that are interesting to read. The key challenge of the news recommendation is to recommend newly published articles. This paper presents a session-based Recurrent Neural Network (RNN) model adapted to the news recommendation. The suggested model uses ConvolutionalNeural Network (CNN) to capture user preferences and adjust referral results [40].

Extractive text allows a set of sentences contained in the real document to be selected and summarized so that the user can better search and grasp the document content. The latest field of research in Extractive summarization uses language modeling (LM) approach. However, the biggest challenge to LM's approach is the way to define sentence representations and accurately determine their parameters. In light of this, this paper examines a new application of the popular Recurrent Neural Network Language Modeling (RNNLM) framework for the extractive summary of news [41].

III. REVIEW SUMMARY

Various approaches to personalized news extraction systems were studied. Table 1 outlines the News extraction systems reviewed in this paper. We have discussed 4 different approaches. Traditional approach fails to capture the underlying semantics of news content. Thus it does not provide a personalized extraction result. Ontology-based systems provide semantically similar news contents as per the user need but do not have a capability of summarizing or generating news from related news stories.NLP tools based systems have limitations in flexibility, the number of rules required, size of domain etc. Neural network based frameworks give a plenty of options and it is yet to be explored. Table 1 recapitulates the systems in the domain of the personalized news extraction framework

Table no 1: Summary of systems in the domain of the personalized news extraction framework

Personalized Extraction Approach	**Reviewed works**	**year**
Traditional Approach (Content Based + Collaborative + Hybrid)	Personal news agent[8]	1999
	Bikini[10]	2001
	Merialdo et al. [11],	1999
	Billsus et al. [12]	2000
	PIN [13]	1998
	PTV [14]	2000
	Google News[9]	2007
	UseNet News [15]	1997
	New Google News recommendation system[16]	2010
Semantic Based (Ontology based)	SmartPush [18]	2001
	SeAN [19]	2001
	aceMedia personalization system[21]	2005
	myPlanet[20]	2001
	SenSee[23]	2007
	Valet et al.[22]	2006
	Hermes framework [26]	2007
	Athena[27]	2010
	ePaper[24]	2009
	News@hand[25]	2008
	ontology-based extractive method for summarization [28]	2010
NLP tools based	data-driven architecture [31]	2017
	TextRank algorithm[32]	2016
	Abstractive Summarization [33]	2017
	sports news, automatic generation systems [34]	2017
Neural Network Based	News storyline generation [38]	2018
	News comments generation [39]	2018
	Session based news recommendation [40]	2017
	RNNLM framework [41]	2015

IV. CONCLUSION

In this paper, we have analyzed and reviewed different approaches to personalized news extraction frameworks. In recent years extensive investigations and analysis have been done in the domain of personalized media extraction. Since neural network based media extraction is a new field with a lot of potentials, it can be expected to grow in different directions. Due to the huge amount of news items produced daily, the public is looking forward to news stories based on related news events which give updates on a particular news topic. Journalist requires an automated tool which can be used in an archival system to extract a news story as per

his/her preference. The use of a neural network in this area is promising and further studies should take place to explore the power of neural networks.

References

[1]. QimingDiao and Jing Jiang. Recurrent chinese restaurant process with a duration-based discount for event identification from twitter: *In Proceedings of the 2014 SIAM International Conference on Data Mining. SIAM*, 2017, 388–397.

[2]. Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang, Timeline generation through evolutionary trans-temporal summarization.: *In Proceedings of the Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics*, 2011, 433–443

[3]. Lifu Huang and Lian'en Huang, Optimized event storyline generation based on mixture-event aspect model. In EMNLP,2013, 726–735

[4]. Jiwei Li and Claire Cardie, Timeline generation Tracking individuals on twitter: *In Proceedings of the 23rd international conference on World wideweb.ACM*,2014, 643–652.

[5]. Ziqiang Cao, Sujian Li, Yang Liu,Wenjie Li, and Heng Ji, A novel neural topic model and its supervised extension*: In Twenty-Ninth AAAI Conference on Artificial Intelligence*,2015,2210–2216.

[6]. Dai H., Mobasher B, Integrating semantic knowledge with web usage mining for personalization, Web Mining: *Applications and Techniques, A. Scime (Ed.), Hershey: Idea Group Publishing*, 2004,276-306

[7]. Balabanovic, M. and Y. Shoham, Fab: *Content-based, collaborative recommendation Communications of the ACM*, 40(3):1997,66-72

[8]. Billsus and Michael J. Pazzani A Personal News Agent that Talks, Learns and Explains,1999

[9]. Abhinandan Das, Mayur Datar, Ashutosh Garg ,Google News Personalization: *Scalable Online Collaborative Filtering,WWW* 2007, May 8–12, 2007

[10]. A.Dorfler, S. Eilert, A. Mentrup, M. E. Muller,R. Rolf, C.R. Rollinger, F. Sievertsen, F. Trenkamp, Bikini: *User Adaptive News Classification in the World Wide Web*,2001

[11]. Bernard Merialdo, Kyung Tak Lee, Dario Luparello, Jeremie Roudaire, Automatic Construction of Personalised TV News Programs, in ACM Multimedia '99, October 99, Orlando, Florida, USA, pp 323-331

[12]. Daniel Billsus, Michael Pazzani, User Modelling for Adaptive News Access, in User Modeling and User Adapted Interaction 10: 2000, 147-180

[13]. Ah-Hwee Tan, Christine Teo, Learning User Profiles for Personalized Information Dissemination,:*In Proceedings of International Joint Conference on Neural Network* 1998, 183-188

[14]. Paul Cotter, Barry Smyth, PTV Intelligent Personalized TV Guides: *In Proceedings of the 17th National Conference on Artificial Intelligence*, AAAI 2000, Austin, Texas, 2000, 957-964

[15]. Joseph Konstan, Bradley Miller, David Maltz, Jonathan Herlocker, Lee Gordon, John Riedl, Applying Collaborative Filtering to UseNet News, in Communications of the ACM March 1997/Vol. 40, No. 3, 77-87

[16]. Jiahui Liu, Peter Dolan, Elin Rønby Pedersen, Personalized News Recommendation Based on Click Behavior, IUI'10, February 7–10, 2010, Hong Kong, China.

[17]. Daya C. Wimalasuriya,,Dejing Dou ,Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches,March 2010

[18]. Sami Jokela, Marko Turnpeinen, Teppo Kurki, Eerika Savia, Reijo Sulonen, The Role of Structured Content in a Personalized News Service, *In Proceedings of the 34th Hawaii International Conference on System Sciences* 2001, pp 1-10

[19]. Liliana Ardissono, Luca Console, Iliara Torre, An Adaptive System for the Personalised Access to News, in AI Communications 14,2001, 129-147

[20]. Yannis Kalfoglou, John Domingue, Enrico Motta, Maria Vargas-Vera, Simon Buckingham Shum,*myPlanet*: an ontology-driven Web-based personalized news service,2001

[21]. David Vallet1, Phivos Mylonas2, Miguel A. Corella1, José M. Fuentes1,Pablo Castells1, and Yannis Avrithis2, A Semantically-enhanced personalization framework for knowledge driven media services,2005

[22]. David Valet and Miriam Fernandez and Pablo Castells and PhivosMylonas and YannisAvrithis "A contextual personalization approach based on ontological knowledge",2006

[23]. Semantics-based Framework for Personalized Access to TV Content: the iFanzy Use Case Pieter Bellekens , Lora Aroyo, Geert-Jan Houben , AnneliesKaptein and Kees van der Sluijs,2007

[24]. BrachaShapira, PeretzShoval, Joachim Meyer, Noam Tractinsky, DuduMimran, ePaper - the Personalized Mobile Newspaper, 2009

[25]. Iván Cantador, Alejandro Bellogín, Pablo Castells , Ontology-based Personalized and Context-aware Recommendations of News Items ,2008

[26]. Borsje, Jethro, Levering, Leonard, Embregts, Hanno and Frasincar, Flavius Frasincar ,Erasmus University Rotterdam, Hermes: an Ontology-Based News Personalization Portal,10 January 2007
[27]. WouterIJntema, Frank Goossen, Flavius Frasincar, Frederik Hogenboom ,Ontology-Based News Recommendation,EDBT 2010, March 22–26, 2010, Lausanne, Switzerland.
[28]. Plaza, Laura &Díaz, Alberto &Gervás, Pablo, Automatic Summarization of News Using WordNet Concept Graphs. IADIS International Journal on Computer Science and Information Systems 5,2010, 45-57.
[29]. Evans, R., Piwek, P., and Cahill, L, What is nlg? http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.8896,2002
[30]. Manishina, E., Lefèvre, F., Besacier, L., Béchet, F., Allauzen, A., and Jabaian, B, *Data-driven natural language generation using statistical machine translation and discriminative learning*, PhD thesis, University of Avignon 2016.
[31]. Leo Lepp¨anen , Myriam Munezero , Mark Granroth-Wilding ,HannuToivonen, Association for Computational Linguistics Data-Driven News Generation for Automated Journalism :*In Proceedings of The 10th International Natural Language Generation conference*: Santiago de Compostela, Spain, September 4-7 2017, 188–197
[32]. Zadbuke et al., *International Journal of Advanced Research in Computer Science and Software Engineering 6 (3),* March- 2016, pp. 124-127
[33]. Riya Jhalani, Yogesh Kumar, Meena, An Abstractive Approach For Text Summarization , *International Journal of Advanced Computational Engineering and Networking*, ISSN: 2320-2106, Volume-5, Issue-1, Jan.-2017
[34]. Gong, W. Ren and P. Zhang, "An automatic generation method of sports news based on knowledge rules," *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, Wuhan, 2017, pp. 499-502
[35]. Eidnes, L. (2015). Auto-generating clickbait with recurrent neural networks https://larseidnes.com/2015/10/13/auto-generatingclickbait- with-recurrent-neural-networks/.
[36]. Karpathy, A., The unreasonable effectiveness of recurrent neural networks, http://karpathy.github.io/2015/05/21/rnn- effectiveness,2015'
[37]. Ayana, Shen SQ, Lin YK et al. Recent advances on neural headline generation, *Journal of Computer Science and Technology* 32(4), July 2017, 768–784
[38]. Zhou, Deyu&Guo, Linsen& He, Yulan, Neural Storyline Extraction Model for Storyline Generation from News Articles. Proceedings of NAACL-HLT 2018, 1727–1736
[39]. H. T. Zheng, W. Wang, W. Chen and A. K. Sangaiah, "Automatic Generation of News Comments Based on Gated Attention Neural Networks," in *IEEE Access*, vol. 6, 2018,702-710
[40]. Park, Keunchan& Lee, Jisoo& Choi, Jaeho, Deep Neural Networks for News Recommendations. 2255-2258. 10.1145/3132847.3133154.Conference: Conference: the 2017 ACM,2017
[41]. K. Y. Chen *et al*., "Extractive Broadcast News Summarization Leveraging Recurrent Neural Network Language Modeling Techniques," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1322-1334, Aug. 2015.

Shine. K. George "Approaches to personalized news extraction framework" IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 7, 2018, pp. 06-11