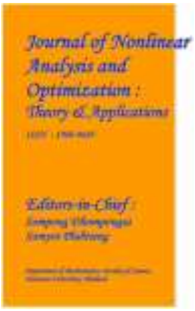
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 2 : 2023
ISSN : **1906-9685**



A FASCINATING AND DYNAMIC ANALYSIS OF A URL BASED WEB SERVICE TO GENERATE INSIGHTFUL THUMBNAILS FOR LINKS

B Lavanya, V Ramesh, G Ravi Kumar, Assistant Professor, Dept. of Computer Science Engineering, Brilliant Institute of Engineering and Technology, Hyderabad, Telangana, India
Muthoju Shreya, Student, Dept. of Computer Science Engineering, Brilliant Institute of Engineering and Technology, Hyderabad, Telangana, India

Abstract

Previews of internet hyperlinks are generally generated primarily based on the metadata captured from the URLcontent. Sometimes, the preview sentencesare extracted by means of content material summarization. Such internet link previews can be visible in distinct apps just like the internet browser, chat app, messaging or e mail apps and many others.These previews are static in nature and do no longer trade with appreciate tochanging context. Therefore, they'll no longer be specifically applicable to the receiver of the link. In this paper, we present a web provider for generating shrewd previews in a talk application, which captures the nearby cause of the consumer from the chat content material and uses it to show most effective applicable content extracted from the previewed URL. Since the consumer rationale can change dynamically, ourmachine generated previews are also dynamic, which alternate on the fly if it detects a change of topic being mentioned within the cutting-edge chat.

Keywords- User intent modelling; web previews; chat application; web service

I. INTRODUCTION

Most mobile applications, including chat, messaging services like WhatsApp,internet browser, web playing cards, socialnetworking apps and many others. Have the ability to generate previews of net links. Such previews make it clean for the user to fast visualize the content of thehyperlink. The internet link preview includes an picture extracted from the URL content material alongside a few text. The text is usually extracted from the URL's metadata. In absence of sufficient metadata, the text can constitute the maximum vital sentences from the article. Web link previews are static, seeing that they are extracted from the internet contentmaterial without thinking about any external context. The extracted statistics shown inside the internet preview mightnot be applicable to the user, if the user is inquisitive about a particular a part of the URL content material. For example, if the user is reading a Wikipedia article on Mexico, the preview might also best provide the internet web page call and few lines related to major topic of the content, while the person can also virtually be interested by Mexican food which is likewise mentioned within the identical page. In this sort of case, it'd be beneficial if the device could infer the subject of the user's hobby or purpose, and display the extracted web content applicable to the subject. Fig. 1 suggests static in addition to dynamic web preview technology for a chat utility on a mobile device. In this paper, we broaden a web carrier for generating dynamic web previews that are relevant to the consumer. Our gadget customizes the internet preview by extracting only facts that the consumer is possibly to be inquisitive about, based at the chat subjects. We anticipate the sort of system will enhance the fine of the consumer revel in and consumer engagement and additionally save the user's time.

We put in force our purpose detection based totally web preview generation service on a chat utility walking on a mobile tool. However, our gadget can in idea be used in any app to generate relevant web link previews. The relaxationof this paper is structured as follows: in the subsequent phase, we survey related paintings inside the place of technology of dynamic previews. Section three offers an overview of our version for reason shooting and preview technology. Section four gives implementation info for a evidence of idea. Section 5 describes a testto discover which sentences

http://doi.org/10.36893/JNAO.2023.V14I2.293-298

customers discover most applicable inside a given URL content, and correlate the user generated outcomes with those generated by means of our algorithm. Section 6 concludes the paper.

II. RELATED WORK

In this segment we survey associated paintings in the vicinity of internet hyperlink previews and their automated generation.

a) Work related to generation of webthumbnails

There are a number of associated works in the area of computerized preview and thumbnail technology. Czervinski [1]studied how web previews should assistusers locate the relevant webpages quicker. Aula [2] as compared the usefulness of textual content and photo primarily based previews and located that a aggregate of both is most useful. Esmaeili [3] discussed a method to generate thumbnails of pictures, the usage of a trained deep neural network to discover a salient area to crop from the authentic image to expose as a thumbnail.

b) Patents related to web previews

A few patents also are to be had that speak techniques to generate net previews. A 1999 IBM patent by means of Wayne Brown [4] proposes a system to generate thumbnail photos of internet pages to show as a search result, where the thumbnail represents how the website would look whilst parsed and opened in an internet browser. Weiss [5] describes a comparable device for parsing and previewing a web site that looks in search engine results. The 2005 Microsoft patent by way of Platt [6] describe an internet link preview machine where the previewed data describes characteristics of the web site inside the link. A Facebook patent [7] mentions an internet preview thumbnail generated upon hovering on a link in an internet browser, wherein the content of the image is a scaled down version of sure capabilities within the webpage. Another Microsoft patent [8] describes an internet preview where the metadata and internet content are summarized to generate an internet preview. However, as mentioned earlier, all of the above related works typically describe static previews, wherein the preview content is extracted from the URLmetadata or content material in the webpage. None of them mention a preview that extracts statistics to display corresponding to the person's interest.

c) Generating more useful previewcontent

Jones [9] defined a surfing utility that extracted and displayed a term cloud of the webpage content, as a extra beneficial answer than regular previews. This work extracted static, if more useful, content from the web site and had no reference to the consumer's present day hobby. Our machine extracts phrases of the consumer's hobby from the current person conduct (inclusive of chat or seek content material) and makes use of these phrases to perceive applicable content to show within the preview. Sarkar et. Al. [14] described a supervised algorithm for contextual summarization of a web site, wherein associated content from links became summarized and proven within the cutting-edge website. Although the dynamically generated precis might be shown as preview inside the surfing scenario, this will no longer be applicable in a usual preview situation e.G. In a talk utility.

III. SYSTEM OVERVIEW

In this section, we describe the numerous modules of our net provider for dynamic preview era for a given URL, interior a talk application on a cellular device. Our system first extracts the subject keywords representing the consumer interest or intent on the time of preview technology. The key phrases are extracted based totally on the encompassing chat logs, with the assumption that the consumer is probably discussing the topic they may be interestedby when the URL link is shared as a part of the chat. These extracted key phrasesare then used to discover which of the sentences from the URL content need to bedisplayed as a part of the preview.

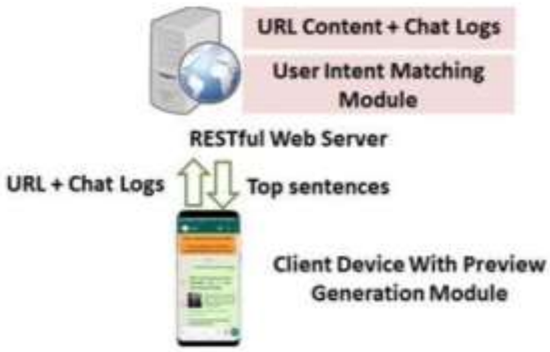

URL Content + Chat Logs
User Intent Matching Module
RESTful Web Server
URL + Chat Logs
Top sentences
Client Device With Preview Generation Module

Figure 2. High level architecture of the web service to generate dynamic user previews on the mobile device.

Our machine is applied as an internet service, wherein the URL is sent to the server along with the extracted keywords or subjects of the consumer's interest, from the chat logs. The server strategiesthe URL and unearths the most relevant sentences from the website content corresponding to the given keywords, which it then returns to the cell tool. All conversation among the server and cell tool happens the use of JSON. Fig. 2 offers a excessive degree architecture diagram of the machine. In the subsequent subsections, we describe every of the additives in element.

A. Keywords Extraction Module: This module is present inside the purchaser device, and captures the key phrases describing the cause. The key phrases are captured from the chat logs after chat segmentation. Since the preview is generated with appreciate to an URL best, it's far vital to pick out the chat messages which relate to a particular URL. In our implementation, we made the subsequent simple assumption: the space between the cutting-edge chat message and previously shared URL is measured with appreciate to

(a) quantity of chat message devices in among the two and (b) time. The closest message is considered to be the only related to an URL. For every such message, we eliminated the forestall phrases and final phrases had been taken into consideration to be key phrases representing the context of chat.

B. Server Communication Module: This module also runs within the patron, and sends the extracted key phrases to the server, together with the URL. The keyphrases are sent the use of REST APIs. Forthe first time, each name and picture are requested. For consequent requests for the same URL, most effective textual contentis requested with respect to changing key phrases.

C. User Intent Matching Module: This module runs on the server, and reveals content matching to the person causerepresented in the form of given keywords.The module takes the URL content as input, plays some preprocessing inclusive of article content material extraction, sentence chunking, putting off prevent phrases and so forth. And then reveals which of the sentences in the URL content material is maximum just like the extracted topics. It ranks the sentences as in line with the similarity, and sends thetop matching sentences again to the customer on the cell device. We used 3 exclusive strategies for ranking sentences that are described in phase 4.

D. Preview Generation Module: This module runs at the RESTful server, and sends the preview sentences again to the consumer device along side their rank and authentic role. The pinnacle ranked sentences from the given URL content material are extracted and sent for preview.

E. Preview UI Rendering Module: This module also runs on the customer tool. It takes the statistics received from the server and generates and displays the preview in the course of the chat. The set of preview sentences are proven of their authentic order to hold coherence and person may additionally make bigger a preview to accommodate more sentences.

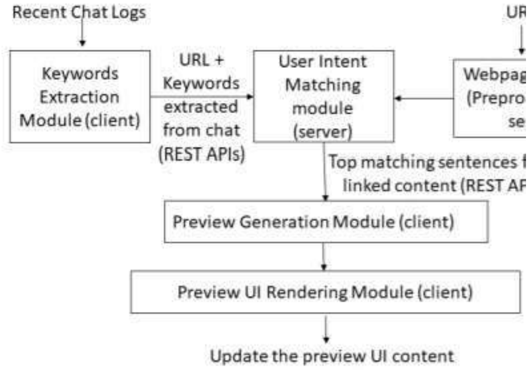

Recent Chat Logs
UR
Keywords Extraction Module (client)
URL + Keywords extracted from chat (REST APIs)
User Intent Matching module (server)
Webpag (Prepro se
Top matching sentences f linked content (REST AF
Preview Generation Module (client)
Preview UI Rendering Module (client)
Update the preview UI content

Figure 3. Flowchart of the steps involved (on the cloud server and client mobile device) to generate dynamic previews

Fig. 3 shows a flowchart of the steps involved in the generation of the intelligentdynamic preview. In the following section we describe the implementation steps inmore detail.

IV. ALGORITHM DETAILS

The steps of the algorithm for generating the dynamic previews are described in the following subsections.

A. Precomputation

We first build a dictionary of English words using English Wikipedia articles by calculating their TF-IDF score. We only keep the top 200,000 words. This is usedto prepare TF-IDF vectors for keywordsset as well as each sentence of the article. We also use these TF-IDF scores tocalculate the weight factor for each of the terms in the keywords set. We use 3 different approaches to measure similarity between the keywords and sentences, but all of them go through the same set ofbasic steps as mentioned in next section.

B. Basic Algorithmic Steps:

1) Preprocessing with keyword extraction. Assign URL to the last chat sentence via chat segmentation method. Remove the stop-words and extract keywords from the last chat sentence. Keyword extraction is done by a simple lookup into a pre-built dictionary. Any word not present in thedictionary is removed. These keywords represent the topic the user is interested in.

2) Preprocessing of the URL content. Extract the article content from the URL's webpage content and chunk the sentences. Convert each sentence into a bag of words after removing stop-words.

3) Determine similarity between sentence and extracted keywords. Calculate a similarity score by computing distance between the set of keywords and each sentence using TFIDF vectors.

4) Sort sentences and display the top matching ones as per the similarity score. Sort the sentences according to the similarity score in descending order. Showpreview with top ranked sentences (we choose 2 by default).

C. TF-IDF and Word2Vec embedding based approaches

We used three different approaches for determining similarity and measuring importance of sentences.

1) Approach 1 – TF-IDF primarily based. This approach converts the keyword set as well as each of the file sentences into TF- IDF vectors. Cosine distance is calculated among the pairs and sentences are ranked in keeping with their distance with the key-word set vector. This method is inspired by using Wan et al.'s [12] Simple-hyperlink approach in which they calculated similarity among anchor sentence and linked document sentences torank them.

2) Approach 2 – Centroid distance withWord2Vec phrase embedding. Thismethod computes the centroid of phrase embedding's for key-word set in addition to bag of words representing each sentenceof the thing. Centroid is computed by taking common of the phrase embedding vectors for each of the phrase. Further cosine distance between the centroids is calculated and sentences are ranked according this distance. We use three hundred dimensional vectors, pre-skilled on Google News information.

Three) Approach 3 – Weighted sum forphrase embedding distances between high-quality-matching phrase-pair. Here, forevery of the phrase from keyword set, we discover the phrase which has the least cosine distance among word embedding vectors for every of the sentences. We takea weighted sum of these distances for every keyword. The weight factor for every of the keywords is calculated using their TF-IDF score.

V. EXPERIMENTAL RESULTS

For our experiment to evaluate our approach for dynamic preview generation, we collected chat logs from a private WhatsApp group having 30 participants for a year since January, 2017. A total of 110 URLs were shared in the group during this period. We removed images, videos, URLs having no article content and all the unrelated chat content which does not have any relation with the shared URLs. After the above preprocessing steps, 54 URLs were selected which had at least one chat message, related to its content. Each URL was mapped with the corresponding chat segment. After this, we asked 2 users to rank the sentences from each article according to the last sequence of chat messages presented for the article. One user belonged to the same group and the other was not part of this group. Ranking was done on a scale of 0 – 2 where 2means most relevant and 0 means not relevant at all. Hence even for the same URL, a different message led to a different ranking. The inter-annotator agreement was measured using Cohen's kappa score. We obtained a score of 0.61, indicating a good agreement. We randomly chose the ranking provided by one of the annotators. We now generated the rankings automatically based on our algorithms and evaluated against the manual ranks.

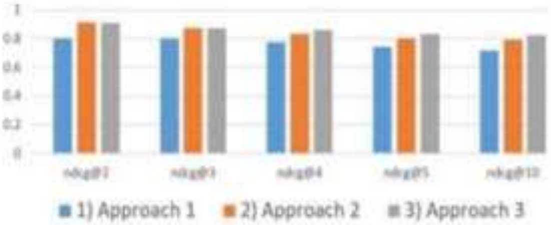


Figure 4. Plot of the NDCG values using each of the three approaches

Fig. 4 shows the comparative results of the normalized discounted cumulative gain (NDCG) where the number of sentences

(n) is varied from 2 to 10. As we can see, the centroid and weight factor approaches perform better than the TF-IDF with cosine similarity approach. This is because the word embedding based methods were able to capture semantic similarity between chat keywords and words from article sentences. Both the previous approaches use word embeddings, but we observed that the centroid based methodperformed better for top ranks while the weighted sum-based method performed better as the number of retrieved sentences increased. In case of the weighed sum approach, the weight factor induced a bias towards more important words from the chat messages, resulting in overall better score. However, its low performance towards top ranked sentences could be dueto the fact that multiple keywords mapped to same word while calculating best matching pair.

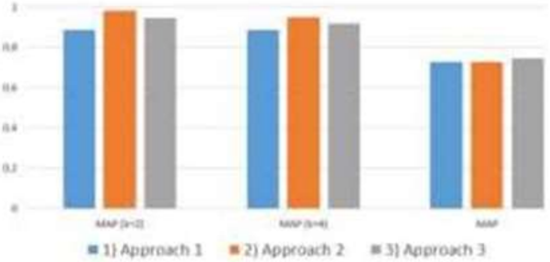


Figure 5. MAP scores for k (sentences to retrieve) set to 2, 4 and total number of sentences in the article

We also made a proof-of-concept prototype creating a sample chat apphaving closed user group of 10 users aged 25-38. In total, 30 URLs were shared within the group and preview was generated using our algorithms. Each ofthe user marked the sentences as relevant

(1) or non-relevant (0). Default previewcontained 2 sentences and the user could expand up to 4 sentences. We calculated mean average precision (MAP) score up tok sentences with k = 2, 4 and number of articles sentences and scores are presented in Fig. 5. These scores are in line with the

previous NDCG values.

CONCLUSION

Present paper described about the details of a prototype net provider implementation, with three techniques for preview generation primarily based on TF- IDF and Word2Vec word embedding. It also provides the gift consequences of an evaluation of the usage of shared URLs from a personal actual-global chat institution in addition to a pattern chat app with a few customers to decide the accuracy of the preview technology machine.

REFERENCES

1. Czervinski, M.P. can Dantzich, M., Robertson, G., and Hoffman, H. The contribution of thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3D. In Proc.INTERACT '99, 163 – 170

2. Anne Aula, Rehan M Khan, Zhiwei Guan, Paul Fontes, and Peter Hong. A comparison of visual and textual page previews in judging the helpfulness of web pages. In Proc. WWW 2010. ACM, 51–60.

3. Seyed A. Esmaeili, Bharat Singh, Larry S. Davis. Fast-At: Fast Automatic Thumbnail Generation Using Deep Neural Networks. in Proc. CVPR 2017.

4. Michael Wayne Brown, Kelvin Roderick Lawrence, Michael A. Paolini. Automatic web page thumbnail generation. US Patent US6356908B1. Filed 1999.

5. Yuval Weiss and Ori Eyal. Systems and methods for generating and providing previews of electronic files such as web files. US patent US7162493B2. Filed 2000

6. John Platt, Ramez Naam, Oliver Hurst-Hiller. Preview information for web-browsing. US patent US20070074125A1. Filed 2005

7. Timothy O'Shaugnessy, Sudheer Agrawal. Presenting image previews of webpages. US patent US9619784B2. Filed 2005.

8. Joseph Masterson, John Gibbon, Eduardo Melo. Inline web previews with dynamic aspect ratios. US Patent US20150278234A1. Filed 2014.

9. Gareth JF Jones and Quixiang Li. Focused browsing: Providing topical feedback for link selectionin hypertext browsing. In Proc. ECIR, 2008. Springer, 700–704.

10. Paige H. Adams, Craig H. Martell, ―Topic Detection and Extraction in Chat,‖ Proc. IEEE ICSC 2008, IEEE Press, Aug. 2008.

11. Han Zhang, Chang-Dong Wang, Jian-Huang Lai, ―Topic Detection in Instant Messages,‖ Proc.ICMLA 2014, IEEE Press, Dec. 2014.

12. Stephen Wan and Cécile Paris. In-browser summarisation: Generating elaborative summaries biased towards the reading context. In Proc. ACL 2008. Association for Computational Linguistics, 129–132.

13. Amit Sarkar, Joy Bose. Methods and systems for generating dynamic previews on electronic devices. India Patent 201841007011. Filed Feb 23, 2018.

14. Amit Sarkar, G. Srinivasaraghavan: Contextual Web Summarization: A Supervised Ranking Approach. In Proc. WWW (Companion Volume) 2018: 105-106.