Temporal Text Summarization of TV Serial Excerpts Using Lingo Clustering and Lucene Summarizer

Sayali Hande, Mrs. M. A. Potey
(Computer Department, DYPCOE, Akurdi, Savitribai Phule Pune University, India)
(Computer Department, DYPCOE, Akurdi, Savitribai Phule Pune University, India)

ABSTRACT: *Text summarization is an art of abstracting key contents from one or more information sources. As time is an important dimension of any information space, it is becoming harder to generate meaningful and timely summaries. While summarizing a story in terms of a timeline, a system may have to extract events and order chronologically. Hence the goal of Temporal Summarization is to develop a system that allows users to efficiently monitor the information associated with an event over time. Previous research algorithms having good speed and scalability share one important shortcoming that, none of them explicitly addresses the problem of cluster description quality. For this reason document clustering is done by using Lingo algorithm in which special emphasis is placed on the quality of cluster labels. Lucene Summarizer is used for text summarization. In many cases users have to spend their maximum time in reading the detail story of entire series of Television (TV) serial episodes which they missed. This paper focuses on a novel application used for automatic generation of meaningful temporal text summarization of missing TV serial excerpts. The user can specify the time period for the content.*

Keywords *- Information Retrieval System, Lingo clustering algorithm, Singular Value Decomposition, Lucene summarizer, Temporal multi-document summarization, Temporal text summarization, Timed Text Rank algorithm .*

I. INTRODUCTION

The size of web is increasing tremendously day by day. To obtain relevant information from the huge data in a limited amount of time is a big challenge. Information Retrieval System (IRS) obtains relevant information from large collection of information resource. IR also support users in browsing or filtering document collections and even does processing a set of retrieved documents. Given a set of documents, clustering is performed to group documents based on the contents. Summarization is the art of abstracting key content from one or more information sources and has become an integral part of everyday life. With the help of summaries people make effective decisions like investments in stock market updates, go to movies on the basis of reviews they have seen. The technology of automatic document summarization is developing and may provide a solution to the information overload problem. Already available tools such as Microsoft's AutoSummarize option in IBM's Intelligent Text Miner, Oracle's Context, and Inxight's Summarizer [1]. It is getting increasingly complex to generate meaningful and timely summaries with the voluminous online data.

Text summarization is the process of automatically creating a compressed version of a given document preserving its information content [2]. There are two types of summarization: extractive and abstractive. Extractive summarization methods simplify the problem of summarization into the problem of selecting a representative subset of the sentences in the original documents. Abstractive summarization may compose novel sentences, unseen in the original sources. The extractive summarization systems are typically based on techniques for sentence extraction and aim to cover the set of sentences that are most important for the overall understanding of a given document. Unsupervised document summarization method is used to create the summary by clustering and extracting sentences from the original document. The summarization process has three phases: analysing the source text, determining its salient points, and synthesizing an appropriate output.

Time and temporal measurements can help recreating a particular historical period or describing the chronological context of a document or a collection of documents [3]. Time plays a central role in any information space, and it has been studied in several areas like information extraction, topic-detection, question-answering, query log analysis and summarization [4]. Basically the goal of the temporal summarization is to develop system that allow users to efficiently monitor the information associated with an event over time. The key problem of summarization is how to identify important content and remove redundant content. Temporal summarization of web pages can be regarded as an extension of content based methods. A page is considered to be a dynamic entity that changes and evolves over

time. To summarize a single web document over a given time interval, first the web page is periodically downloaded with a certain frequency over that interval. The temporal versions of the page are then analysed and any changes are identified. These changes are extracted and used to construct a summary. Ruifang He *et.al.*[5] presents Temporal multi-document summarization (TMDS) as natural extension of multi-document summarization, which captures evolving information of a single topic over time. TMDS mines the temporal characteristics at different levels of topical detail in order to provide the cue for extracting the important content.

In today's busy world people either does not have sufficient time or does not wish to spend the time reading the entire story of TV serial excerpts which they missed. Often they are interested in knowing the updates of episodes without spending much time. They can get the summary of episodes by using this system. Thus, our main objective is to provide users (e.g., specially a daily soap viewers) the compressed version of a detailed story of TV serial excerpts preserving its information content by using temporal text summarization approach. The idea behind our system is to quickly find the summary of storyline in specified time range so as to help the user in understanding the story of missed episodes by using Lingo clustering algorithm and Lucene summarizer.

II. LITERATURE SURVEY

2.1 Event Based Temporal Summarization

Event-based summarization extracts and organizes summary sentences in terms of the events elaborated by the sentences. By connecting event terms with semantic relations, Maofu Liu *et.al.[*6] build up event term graph where relevant terms are grouped into clusters assuming each cluster represents a topic of documents. They investigated two summarization strategies by selecting one term as the representative of each topic so as to cover all the topics, or selecting all terms in one most significant topic so as to highlight the relevant information related to this topic. The selected terms are then responsible to pick out the most appropriate sentences describing them. The evaluation of clustering-based summarization on Document Understanding Conference (DUC 2001) document sets shows encouraging improvement over the well-known PageRank-based summarization.

2.2 TSCAN

TSCAN (A Content Anatomy Approach to Temporal Topic Summarization); a topic is defined as a seminal event or 2 activity along with all directly related events and activities. It is represented by a chronological sequence of documents published by different authors on the Internet. C.C. and M.C. Chen [7] defined a task called topic anatomy, which summarizes and associates the core parts of a topic temporally so that readers can understand the content easily. The proposed topic anatomy model, called TSCAN, derives the major themes of a topic from the eigenvectors of a temporal block association matrix. Then, the significant events of the themes and their summaries are extracted by examining the constitution of the eigenvectors.

Finally, the extracted events are associated through their temporal closeness and context similarity to form an evolution graph of the topic. Experiments based on the official TDT4 corpus [8] demonstrate that the generated temporal summaries present the storylines of topics in a comprehensible form. In terms of content coverage, coherence, and consistency, the summaries are superior to those derived by existing summarization methods based on human composed reference summaries.

2.3 Theme Based Summarization

Multi-document summarization aims to produce a concise summary that contains salient information from a set of source documents [9]. Since documents often cover a number of topical themes with each theme represented by a cluster of highly related sentences, sentence clustering plays a pivotal role in theme-based summarization. Real world datasets always contain noises which inevitably degrade the clustering performance, we incorporate noise detection with spectral clustering to generate ordinary sentence clusters and one noise sentence cluster. By making the theme-based summaries biased towards a users query. Zhang *et.al.*[10] have demonstrated the effectiveness of proposed approaches by both the cluster quality analysis and the summarization evaluation conducted on the DUC generic and query oriented summarization datasets.

2.4 Temporal Event Clustering

Gung and Kalita [11] investigate the use of temporal information for improving extractive summarization of historical articles. Their method clusters sentences based on their timestamps and temporal similarity. Each resulting cluster is assigned an importance score which can then be used as a weight in traditional sentence ranking techniques. Temporal importance weighting occurs consistent improvements over -baseline systems.

2.5 Timed Text Rank Algorithm

Graph-ranking based algorithms (e.g. TextRank) have been proposed for multi-document summarization in recent years. However, these algorithms miss an important dimension, the temporal dimension, for summarizing evolving

topics. For an evolving topic, recent documents are usually more important than earlier documents because recent documents contain much more novel information than earlier documents and a novelty-oriented summary should be more appropriate to reflect the changing topic. Wan, X. [12] propose the TimedTextRank algorithm to make use of the temporal information of documents based on the graph-ranking based algorithm. A preliminary study is performed to demonstrate the effectiveness of the proposed TimedTextRank algorithm for dynamic multi-document summarization. It adds the Temporal Dimension to Multi-Document Summarization. This existing TimedTextRank algorithm focuses only on recent documents but temporal summarization focuses earlier documents as well as recent documents.

2.6 Event Related Updates in Wikipedia

Wikipedia Event Reporter, a web-based system that supports the entity-centric, temporal analytics of event-related information in Wikipedia by analysing the whole history of article updates was described by M. Georgescu *et.al*.[13]. The system first identifies peaks of update activities for the entity using burst detection and automatically extracts event-related updates using a machine-learning approach. Further, the system deter- mines distinct events through the clustering of updates by exploiting different types of information such as update time, textual similarity, and the position of the updates within an article. Finally, the system generates the meaningful temporal summarization of event-related updates and automatically annotates the identified events in a timeline.

2.7 Lingo Clustering

Many approaches to search results clustering have been proposed, including Suffix Tree Clustering (STC), Semantic On-line Hierarchical Clustering (SHOC) and Tolerance Rough Set Clustering (TRC). With their respective advantages such as speed and scalability, all these algorithms share one important shortcoming: none of them explicitly addresses the problem of cluster description quality. This, leads these algorithms to knowing that certain documents should form a group and at the same time being unable to concisely explain to the user what the group's documents have in common. S. Osinski [14] proposed an algorithm called Lingo [15] in which special emphasis was places on the quality of cluster labels. Lingo algorithm reverses the usual order of the clustering process. It first identified meaningful cluster labels using the Singular Value Decomposition (SVD) factorisation, and only then assigned documents to these labels to form proper clusters. For this reason this algorithm could be considered as an example of a description-comes-first approach.

They have compared Lingo with two other algorithms STC and TRC designed specifically for clustering of search results and performed their experiments using data drawn from a large human-edited directory of web page summaries called Open Directory Project. It states that the description-comes-first approach to search results clustering implemented by Lingo significantly outperformed both STC and TRC in topic separation and outlier detection tests.

III. IMPLEMENTATION DETAILS

In today's busy world people watch certain TV serials regularly especially the daily soap viewers may miss few episodes due to some reason. They can get a summary of those missed episodes by using this proposed system. User has to specify start date as well as end date of TV show episode so that the system will generate summary of specified time period. Lucene Summarizer is used to summarize the document.

3.1 System Architecture

The architecture of proposed system is depicted in Fig. 1. User sends request to TV serial server for extracting the summary of TV serial excerpts. User can specify particular time period by selecting a start and end dates of TV serial episodes with respect to day, month and year from a given calender. To extract the summary of specified time period, TV serial server applies pre-processing steps on the selected date's documents. Pre-processing steps such as stemming, stop word removal are applied on the input documents. Input documents are nothing but detailed textual story of a TV serial.

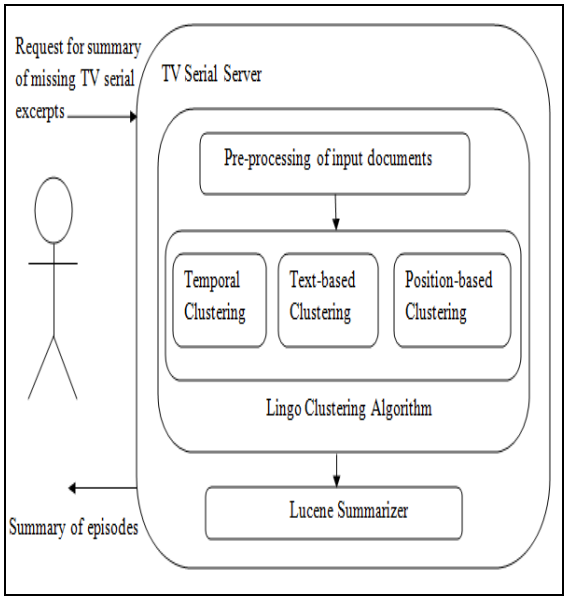

Fig. 1 Proposed System Architecture

The three approaches of clustering are applied as follows:

1) Temporal clustering uses burst detection algorithm presented in [16] to identify bursts among the TV serial excerpts.
2) Text based clustering is an incremental clustering based on Jaccard similarity as a distance measure.
3) Position based clustering assumes that sentences on the same topic are located in spatial proximity of each other on the excerpt page, by investigating the positions of all sentences modified in a burst can identify position clusters [17].

Labeling of cluster is done by using Lingo clustering algorithm which helps to generate text classification. Lucene summarizer [18] tokenizes the input into paragraphs, and the paragraphs into sentences. Then writes each sentence out to an in-memory Lucene index for the document with sentences as fields in single-field Lucene documents with index time boosts specified by the paragraph and sentence number. It then computes the term frequency map of the index to find the most frequent words found in the document, takes the top few terms and hits the index with a Boolean OR query to find the most relevant sentences [19]. It return the top few sentences ordered by Lucene document id thus found constitute the summary. Then it gives summarized result to the user.

3.2 *Algorithm*

We have used Lingo algorithm [15] for clustering of documents. It first identifies meaningful cluster labels and then only assigns search results to these labels to build proper clusters. Our proposed system implements following steps[14]:

1) Preprocessing of the input snippets of TV serial excerpts, which includes tokenization, stemming and stop-word marking.

2) Identifies words and sequences of words frequently appearing in the input documents.

3) A matrix factorization is used to induce cluster labels for TV serial excerpts.

4) Snippets are assigned to each of these labels to form proper clusters. The assignment is based on the Vector Space Model (VSM) and the cosine similarity between vectors representing the label and the snippets.

5) Postprocessing, this includes cluster merging and pruning.

3.3 Mathematical Model

IRS has three components: input, processor and output. Proposed system contains TV serial excerpts as an input and the processor performs operations such as pre-processing, frequent phrase extraction, cluster label induction, cluster content discovery, cluster formation, summarization in response to a query. Finally, we came up with summarized result as an output.

1) Preprocessing

$D = \{d1,d2,..,dn\}$ . . . . . (1)

For all $d \epsilon D$ perform text segmentation of d means detecting word boundaries etc. If language of d recognized then apply stemming and mark stop-words in d.

2) Frequent Phrase Extraction

Concatenate all documents.

$P_c \leftarrow$ discover complete phrases.

$P_f \leftarrow p : \{p \epsilon P_c \wedge$ *frequency(p)* > TermFrequencyThreshold$\}$

3) Cluster Label Induction

A ←term-document matrix of terms not marked as stop-words and with frequency higher than the Term Frequency Threshold

$\Sigma,U,V \leftarrow$ Clustering(A);

{Product of Clustering of A}

$k \leftarrow 0$; {Start with 0 clusters}

$n \leftarrow$ rank(A);

repeat

$k \leftarrow k + 1$;

$q \leftarrow (\Sigma^{k}_{i=1} \Sigma_{ii}) / (\Sigma^{n}_{i=1} \Sigma_{ii})$ . . . . . (2)

until $q <$ Candidate Label Threshold ;

$P \leftarrow$ phrase matrix for P_f ; for all columns of $U^{T}_{k} P$

do

find the largest component m_i in the column;

add the corresponding phrase to the Cluster Label

Candidates set;

labelScore $\leftarrow m_i$;

calculate cosine similarities between all pairs of candidate labels;

identify groups of labels that exceed the Label SimilarityThreshold;

for all groups of similar labels do

select one label with the highest score;

4) Cluster Content Discovery

for all $L \epsilon$ Cluster Label Candidates do

create cluster C described with L;

add to C all documents whose similarity to C exceeds the Snippet Assignment Theshold;

put all unassigned documents in the "Others" group;

5) Cluster Formation

for all clusters do

clusterScore ← labelScore×‖ C ‖; . . . . . (3)

6) Summarization

Maximum Marginal Relevance (MMR) method aims to reduce redundancy in summaries.

$Score_s = \lambda Sim(S;Q) - (1-\lambda)Sim(S,Sum)$ (4)

λ = weight factor;

S = a sentence which is not yet in the summary;

Sum = summary;

Q = hidden query;

Sim(S,Q) = a score representing how good the sentence is to be used in a summary (without taking redundancy into account);

The second part of the formula subtracts score if the sentence is too similar to the existing summary.

3.3 Dataset

For better results of temporal text summarization of TV serial excerpts we have crawled 500 documents. Our proposed system uses real time dataset named as desitvforum.net [20] for crawling any TV serial's episodic textual detailed story. We have crawled 101 detailed episodes of Star Plus channel's daily soap "Everest". We can crawl more than one TV serial's excerpts and it will perform same operations as above.

IV. RESULTS AND DISCUSSION

Fig. 2 shows graph of Accuracy value for Lingo clustering algorithm. Cluster1 indicates Temporal Clustering, Cluster2 indicates Text-based Clustering and Cluster3 indicates Position-based Clustering technique. Accuracy value of Cluster2 i.e. Text-based Clustering is more than Cluster1 and Cluster3.

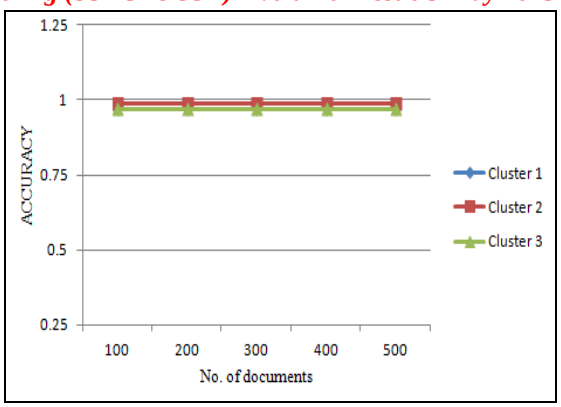

Fig. 2 Graph of accuracy showing results over 500 documents

Following Table I displays accuracy results of Lingo clustering algorithm considering three types of clusters.

TABLE I
ACCURACY

No. of documents	Cluster 1	Cluster 2	Cluster 3
100	0.97	0.99	0.97
200	0.97	0.99	0.97
300	0.97	0.99	0.97
400	0.97	0.99	0.97
500	0.97	0.99	0.97

Fig. 3 shows graph of Precision value for Lingo clustering algorithm. This graph depicts that the Cluster 2 i.e. Text based clustering method is more efficient than other two clustering methods.

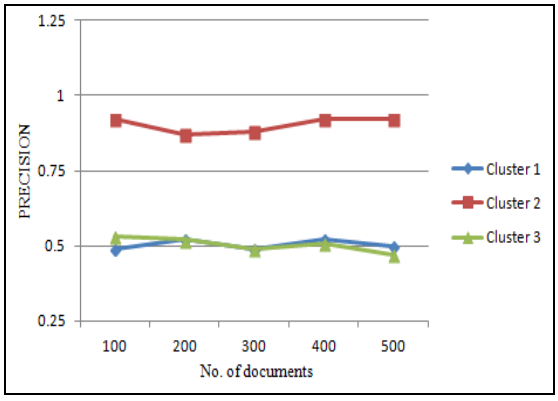

Fig. 3 Graph of precision showing results over 500 documents

Following Table II shows precision results of Lingo clustering algorithm using three types of clusters considered.

TABLE III
PRECISION

No. of documents	Cluster 1	Cluster 2	Cluster 3
100	0.49	0.92	0.53
200	0.52	0.87	0.52
300	0.49	0.88	0.49
400	0.52	0.92	0.51
500	0.50	0.92	0.47

Fig. 4 shows graph of Recall value for Lingo clustering algorithm.

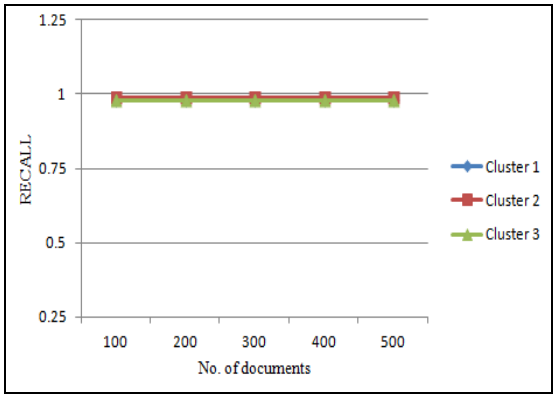

Fig. 4 Graph of recall showing results over 500 documents

Following Table III depicts recall results of Lingo clustering algorithm considering three types of clusters.

TABLE IIIII
RECALL

No. of documents	Cluster 1	Cluster 2	Cluster 3
100	0.98	0.99	0.98
200	0.98	0.99	0.98
300	0.98	0.99	0.98
400	0.98	0.99	0.98
500	0.98	0.99	0.98

V. CONCLUSION

Temporal Summarization is used to develop a system that allows users to regularly monitor the information associated with an event over time. Our proposed system acquiesce users to smoothly extract the summary of TV serial excerpts by using temporal text summarization technique. Various approaches have been used such as temporal event clustering, text-based clustering and position-based clustering for performing temporal summarization. Document clustering is done by using Lingo algorithm and Lucene summarization algorithm is used for text summarization. A user can get crisp and date wise sequential story of missing TV serial excerpts after the selection of dates from the calendar. In this way system reduces the time required to read the complete episode stories.

VI. FUTURE WORK

News reporters can use this system for creating headlines or they can convert lengthy interviews into short summary. Our proposed system can be used for extracting temporal textual summary from episodic videos like triology by using Internet Movie Database(IMDB) for movie story, Massive Open Online Course(MOOCS), talkseries(TV series), Web based seminar(webinar), etc. Also we can create similar system for mobile application. In future we can perform video or audio clips summarization using image processing technique remain for future work.

VII. Acknowledgements

We express our sincere gratitude to publishers, researchers for making their resource available & teachers for their guidance. We also thank the college authority for providing the required infrastructure and support.

REFERENCES

[1] U. Hahn and I. Mani, The challenges of automatic summarization, *IEEE Computer*, *Vol. 33, No. 11*, 2000, pp. 29–36.

[2] Alguliev, Rasim, and Ramiz Aliguliyev. "Evolutionary algorithm for extractive text summarization." *Intelligent Information Management 1.02 (*2009): 128.

[3] O. Alonso, J. Strotgen, R. A. Baeza-Yates, and M. Gertz, Temporal information retrieval: Challenges and opportunities, *TWAW*, *vol. 11*, pp. 1-8, 2011.

[4] I. Mani, J. Pustejovsky, and B. Sundheim, Introduction to the special issue on temporal information processing, *ACM Transactions on Asian Language Information Processing (TALIP), vol. 3, no. 1*, pp. 1-10, 2004.

[5] R. He, B. Qin, T. Liu, and S. Li, Cascaded regression analysis based temporal multi-document summarization, *Informatica: An International Journal of Computing and Informatics, vol. 34, no. 1*, 0 pp. 119-124, 2010.

[6] M. W. Q. L. Maofu Liu1, Wenjie Li, Extractive summarization based on event term clustering, *Proceedings of the ACL 2007 Demo and Poster Sessions*, p. 185-188, *Association for Computational Linguistics*, June 2007.

[7] C. C. Chen and M. C. Chen, Tscan: A content anatomy approach to temporal topic summarization, *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 1, pp. 170-183, 2012.

[8] http://www.itl.nist.gov/iad/mig/tests/tdt/2003/papers/ldc.ppt, 2003.

[9] A. Jatowt and M. Ishizuka, Temporal multi-page summarization, *Web Intelligence and Agent Systems*, vol. 4, no. 2, pp. 163-180, 2006.

[10] D. G. W. L. Xiaoyan Cai, Renxian Zhang, Simultaneous clustering and noise detection for theme-based

summarization, *Proceedings of the 5th International Joint Conference on Natural Language Processing*, p. 491-499,November 2011.

[11] J. Gung and J. Kalita, Summarization of historical articles using temporal event clustering, *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 631-635, Association for Computational Linguistics, 2012.

[12] X. Wan, Timedtextrank: adding the temporal dimension to multi-document summarization, *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 867-868, ACM, 2007.

[13] M. Georgescu, D. D. Pham, N. Kanhabua, S. Zerr, S. Siersdorfer, and W. Nejdl, Temporal summarization of event-related updates in wikipedia, *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 281-284, International World Wide Web Conferences Steering Committee, 2013.

[14] S. Osinski, Improving quality of search results clustering with approximate matrix factorisations, *Advances in Information Retrieval*,pp.167-178, Springer, 2006.

[15] Osiński, Stanisław, Jerzy Stefanowski, and Dawid Weiss, Lingo: Search results clustering algorithm based on singular value decomposition, *Intelligent information processing and web mining. Springer Berlin Heidelberg*, 2004. 359-368.

[16] Y. Zhu and D. Shasha., Efficient elastic burst detection in data streams, *Proceedings of KDD* '03, 2003.

[17] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, A study on position information in document summarization, *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 919-927, Association for Computational Linguistics, 2010.

[18] Hägerstrand, Anton, *Multi Document Summarization*, School of Computer Science and Engineering Royal Institute of Technology , 2011.

[19] Kelly, Liadh, et al. "Report on summarization techniques." Khresmoi project deliverable D 4 (2013): 4. Hanbury, Allan. "Medical information retrieval: an instance of domain-specific search." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[20] *DesiTVForum* 2015. DesiTVForum. Available at

http://desitvforum.net/television/everest/