A LATTICE-BASED APPROACH FOR NEWS CHAIN CONSTRUCTION

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Computer Engineering

by
Mustafa TOPRAK

July 2015
İZMİR

We approve the thesis of **Mustafa TOPRAK**

Examining Committee Members:

Asst. Prof. Dr. Selma TEKİR
Department of Computer Engineering, İzmir Institute of Technology

---

Asst. Prof. Dr. Belgin ERGENÇ BOSTANOĞLU
Department of Computer Engineering, İzmir Institute of Technology

---

Asst. Prof. Dr. Mutlu BEYAZIT
Department of Computer Engineering, Yaşar University

14 July 2015

Asst. Prof. Dr. Selma TEKİR
Supervisor, Department of Computer Engineering
İzmir Institute of Technology

Prof. Dr. Halis PÜSKÜLCÜ
Head of the Department of
Computer Engineering

Prof. Dr. Bilge KARAÇALI
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Selma TEKİR who suggested a project that would help me improve my background on the major I like to study. Moreover, I am grateful for her patience and effort during both research and development phase and for the chance to benefit from her knowledge as a guide.

I thank my co-advisor Jens ALLMER for sharing his scientific interpretation perspective on different domains.

I would like to thank Belgin ERGENÇ BOSTANOĞLU and Mutlu BEYAZIT for accepting to be a member of my thesis defense jury and spending their time to evaluate my study.

Of course, I should thank a lot my friends for all the support. They were one of my motivations during stressful and tiring times.

Finally I thank my family for creating physical and mental conditions for my education during my whole life. They are always trying to make my life more comfortable without any expectations. Any of my projects will have no meaning without them.

ABSTRACT

A LATTICE-BASED APPROACH FOR NEWS CHAIN CONSTRUCTION

Each news article and column can be part of a manually created news story or chain by journalists and columnists. However, increasing amounts of data published by news companies each year makes manual analysis thus creation of news stories and chains almost impossible. When the amount of data is considered, it is obvious that automated systems' support is vital to journalists, columnists and intelligence analysts.

A news chain is a set of news articles that form a connected and coherent whole. In the traditional "connecting the dots" approach, news chains are constructed based on given two articles as start and end news of the chain. In this study, a method is proposed to create coherent news chains without the predetermination of start and end articles of the chain. Intuition of the method comes from the partial order relation among news articles. We try to show that lattice structure can represent relation or hierarchy among news articles that have a partial order in nature. Creating concept lattice is prepared out of the inverted index structure of news articles which is one of the main contributions of the study.

In the experimental work, an artificial dataset is processed to show the steps of the method. After that, we also provide the evaluation using real dataset results.

ÖZET

HABER ZİNCİRİ OLUŞTURULMASI İÇİN KAFES TABANLI BİR YAKLAŞIM

Haber kaynaklarından okunan her bir haber veya köşe yazısı aslında yazarlar ya da gazeteciler tarafından elle oluşturulmuş bir haber hikayesi veya zincirinin bir halkası olarak kabul edilebilir. Ancak, her yıl haber kaynakları tarafından yayınlanan ve gittikçe artan haber sayısı göz önüne alındığında, haberlerin elle analizi gittikçe zorlaşmakta ve hatta imkansız hale gelmektedir. Yayınlanan haber sayılarındaki artış ve oluşan haber arşivlerinin büyüklüğü değerlendirildiğinde, haber analizindeki otomasyon sistemleri, özellikle araştırmacı gazetecilik ve istihbarat analizi alanlarında vazgeçilmez bir ihtiyaç haline gelmektedir.

Haber zinciri, kendi aralarında anlamlı ve bağlı bir bütün oluşturan haber dizisidir. Geleneksel "noktaları birleştirelim" yaklaşımında haber zincirleri iki uç noktasının arası doldurularak oluşturulmaktadır. Bu çalışma kapsamında, başlangıç ve bitiş haberleri önceden belirlenmemiş, tutarlı bir haber zinciri oluşturan bir metod önerilmektedir. Bu metodun dayandığı fikir, haberler arasındaki kısmi sıralama (partial order) bağıntısıdır. Kafes yapısının (lattice) haberler arasındaki ilişkiyi ve hiyerarşiyi modelleyebileceği ve temsil edebileceği üzerinde durulmaktadır. Kafes yapısının oluşturulmasında girdi olarak ters dizin (inverted index) yapısının kullanılması çalışmanın bir başka özgün noktasıdır.

Deneysel çalışmalar bölümünde yapay bir veri seti üzerinde önerilen metodun aşamaları gösterilmektedir. Ardından gerçek veri setleri üzerinde elde edilen sonuçlar yorumlanarak sunulmaktadır.

TABLE OF CONTENTS

APPENDICES

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

LIST OF ABBREVIATIONS

IR	Information Retrieval
FCA	Formal Concept Analysis
TF	Term frequency
IDF	Inverse Document Frequency
POSET	Partially Ordered Set
XML	Extensible Markup Language

CHAPTER 1

INTRODUCTION

A news chain is an ordered, not essentially fully ordered, set of news articles where links representing the relation between articles make logical sense. There can be different aspects for the order of articles; such as time, event, and subject, time being the most common one.

News articles generally are considered as dots, discrete points, if no relation is declared between them. For creation of a news chain, pairwise relations should be established between articles (dots). Given two endpoints, finding out a sequence of articles with consistently strong pairwise relations between them is called "connecting the dots" (Shahaf and Guestrin, 2010).

During the news chain creation process, existence of coherence through the chain has a crucial importance. Each ordered set of news articles may not have coherence. Therefore, coherent news chain creation ensures consistent relation between consecutive articles and smooth topic transition. This point is directly related with semantic analysis. To have a coherent news chain means semantically-related articles and semantic transitions between the chain points.

When reading a news article or column, actually we are reading a news chain (generally coherent according to writer) rendered by journalists or columnists. General approach for creation of a news story or chain published in newspapers is manual creation. Within the limits of journalists' memory and research, relations between news articles (in the content of articles, we can call it subjects) are created and some of the information is rendered by journalist and finally created news chain is presented to the readers. However, when the amount of news articles is considered, complication of manual news chain creation can be understood and thus correctly judged as hard and even impossible. For years, lots of news articles are published by lots of news companies. Therefore, archives of news companies are expanding day by day. Even when we consider the corpora of just two companies like Reuters and The New York Times, The Reuters Corpus Volume 1 has approximately 808.000 news articles and The New York Times Annotated Corpus has approximately 80.000 news articles. These numbers shows how hard it is to construct news chains manually.

Relating news, discovering news stories have become vital for all of us. However, some majors like investigative journalism and intelligence analysis require making connections in the flow of news as part of their daily work and need software tool support. Particularly methods for automated news chain construction have become an important necessity for these professions.

The aim of this study is to develop a method for automatic construction of news chains. When we consider the concept of news chain, time-ordered set of articles is generally assumed. But, time restriction by itself is not sufficient as a strict order; we have wide time ranges and seek consumable size of chains, thus it is also necessary to create a semantic hierarchy between the news articles.

When the general concept of a relation between the news articles is considered, we can say that it is possible to partially, not full/strict, order the given news articles. For example, assume that we have three articles, the first document is related with “drug”, the second one is about “drug and school” and the third one is on the topic of “drug and its effect on national finances”. We can set up a time-order relation in a straightforward manner, but when semantically considered, the first article will be at the bottom of the hierarchy because of its generality and the other two documents will be at the same level with each other without a direct relation but being linked through the first document in the hierarchy.

The intuition of this study is based on exploration of partial order relations between news articles. Creation of hierarchy between news articles can provide us a pool of news chain candidates. Creation of partially ordered news sets based on the content is the key point of the project. Then, through the use of pruning (by the general lattice pruning criteria and a heuristic measuring how good (coherent) a news chain is) on this hierarchy, useful news chains will be extracted.

CHAPTER 2

BACKGROUND

In this section, description of methodologies that are used in the study, will be presented. The general topics of the utilized methods can be classified as Information Retrieval (IR) methods and Formal Concept Analysis (FCA) methods.

2.1. Information Retrieval Methods

The subject of information retrieval works on constructing a system to extract information from data, that can be text, video, audio or other materials, towards the user needs (Kowalski, 1997). The success of an information retrieval system is assigned based on how it can minimize (compress) the information, the amount of related information extracted for user needs and how fast it retrieves information before presenting to user. Ancestor of information retrieval systems can be accepted as hard-copy catalogues, central repositories (Kowalski, 1997). To access information in an efficient way, catalogues are created such as telephone catalogues of cities, book catalogues of libraries.

After computers have become easily accessible as commercial and individual devices, information storage in the electronic environment has increased and the same problem, accessing target information in big data stores have become a common problem of this new environment, too.

In this section, some of the information retrieval methods that are used will be explained.

2.1.1. Stop Words Removal

Stop words are common words which generally occur in all documents, therefore in most cases they do not imply specific semantic meaning in document content (Manning et al., 2008). As an example, word “of” possibly occurs in the most of the documents and does not have an important weight for the text processing if the sentence it occurs in is “house of him”, but in the example “President of the United States” (Manning et al., 2008), it adds an extra meaning by specifying being president of a specific country.

One of the advantages of stop words removal method is to reduce size of index of documents (Manning et al., 2008). Reducing index size also leads to reduce the computation time of method.

Stop word removal is generally applied by using manually created stop words list. If any of the document words is matched with one of the stop words from the stop words list, then that word is removed from the document. There are different stop words lists released by different research groups based on different aims and understanding.

2.1.2. Stemming

Stemming is a grammatical modification of a word by using some defined rules, generally by chopping off the end of the word (Manning et al., 2008). This is used for overcoming different suffixes of the same word. For example, controls and controlling have the common stem control, therefore instead of using raw version of these words as separate words, depending on the aim, these two words can be accepted as the same word.

Stemming is generally referred to as a heuristic approach (Manning et al., 2008), meaning that it does not guarantee to find exact and morphologically true stem of the word. Stemming approaches generally consist of discrete defined rules, and when these rules are applied to words, which have different meaning, can be stemmed to the same word. For example, if Porter Stemmer (Porter, 1997), a specific stemming algorithm, is used, then words “universal”, “universe” and “university” which implies different meaning stemmed to the same word which is “univers” even though they are expected to have different stems since they imply different meanings.

2.1.3. Inverted Index/File

According to the definition of Donald Knuth, Inverted Index/File is a technique to retrieve secondary key of the data (Knuth, 1998). For example, if the primary structure of a dictionary is Russian-English, than inverted file/index of this dictionary is English-Russian. This structure fastens the search of secondary keys.

In the scope of information retrieval or text mining, inverted index points the inverse relation between document and words. In the object relation, document consists of words. Therefore, we generally say that documentA contains words word1, word2, word3, ... Depending on the topic of the study, we may need to interrogate the secondary keys which are generally words. To fasten the process of finding documents in which word1 occurs, inverted index is preferred.

2.1.4. TF - IDF (Term frequency - Inverse Document Frequency)

Each text document can be defined by each word that is contained in it. According to boolean approach, it is acceptable to say each word has the same importance to define a document regardless of how many times it occurs in the document. Thus, occurring in the document is accepted as 1 and not occurring is accepted as 0 in boolean retrieval.

In real world problems, boolean approach is not always suitable for the representation of the problem. Number of occurrences of the word may provide us with additional information about the document. For example, in the documents on the topic of politics, we expect words, which define the topic or related with the topic, to occur more than the words related to sports. Frequency of word “president” is expected to be greater than the frequency of word “tennis” in the politics domain. Therefore, to weight words for importance in or relation with the document, term frequency (occurrence number in the document) is used as an alternative method. Term frequency is not always used as a raw number, different modifications can be applied, such as normalizations or transformations like TF-IDF.

In the term frequency method, each word in the collection is assumed to have equal importance (Manning et al., 2008). To discriminate the importance of a word in the collection, to overcome the problem of equal importance, document frequency of the term is used (Manning et al., 2008). Document frequency refers to the number of documents in the collection that the term occurs in. Inverted document frequency is calculated as:

$$idf_t = \log \frac{N}{df_t} \tag{2.1}$$

where idf_t is inverted document frequency of term t, N is total number of documents exist in dataset and df_t is number of documents term occurs in.

$tf - idf$ is a weighting method calculated with the combination of tf and idf mentioned above. By this weighting method, instead of using just raw frequency of a term in a document, the number of documents that the term occurs in is used as well. Therefore, composite weight is produced as a result to discriminate terms' importance in the collection. $tf - idf$ value can be calculated as:

$$(tf - idf)_{t,d} = tf_{t,d} \times idf_t \tag{2.2}$$

where $(tf - idf)_{t,d}$ is $tf - idf$ weight of term t in a document d, $tf_{t,d}$ is term frequency of term t in the document d and idf_t is inverted document frequency of term t.

There can be three interpretations of tf-idf weight (Manning et al., 2008):

- TF-IDF value is high if term occurs frequently in small subset of documents (we can interpret that the term is a specific word for related documents)
- TF-IDF value is low if term occurs in most of the documents (we can interpret that term can be assumed as a stop word or regular/common word for the document set)
- TF-IDF value is close to the average value if term occurs in small amount of documents a few times or occurs in almost an average number of documents.

2.2. Lattice Theory and Formal Concept Analysis Methods

Method we suggested for this study is based on the intuition of exploring partial order relations between news articles. Therefore, creation of lattice by using inverted index of news articles is main focus of the study. By using selected terms, which are used to create themes of news chains around them, as extents and news articles (or documents) as intents, concepts are produced and hierarchy between them are structured.

In this section, definitions and some properties of partial order, lattice theory and formal concept analysis are given to explain method suggested in the scope of this study. Since lattice structure is a partially ordered set, section starts with explanation of partial order. Then, by using partial order definition, lattice theory is explained. Finally, Formal concept analysis and its relation with lattice theory is described.

2.2.1. Lattice Theory

Partial Order is a binary relation $\leq$ on a set X that satisfy three properties mentioned below (Border, 2011):

1. Reflexivity:

$$x \leq x, \text{ for all } x \epsilon X \tag{2.3}$$

2. Transitivity:

$$\text{if } x \leq y \text{ and } y \leq z, \text{ then } x \leq z \text{ for all } x, y, z \epsilon X \tag{2.4}$$

3. Antisymmetry:

$$\text{if } x \leq y \text{ and } y \leq x, \text{ then } x = y \text{ for all } x, y \epsilon X \tag{2.5}$$

A **partially ordered set** (or **poset**) is a set, let say X, that is not an empty set and satisfy the three properties mentioned above for each element. A poset can be represented as $(X, \leq)$ (Gallier, 2014). If every element pair is ordered in the set (called **complete**), then partial order is called a **linear order** (or **total order** (Gallier, 2014)) (Border, 2011). For example, real numbers is a linear ordered set. A subset, that is ordered linearly, is called a **chain**.

Strict order, $<$, is the relation defined as $x < y$ *iff* $x \leq y$ and $x \neq y$, then it means the set does not satisfy reflexivity (Gallier, 2014).

Let A be a subset of X where X is a poset $(X, \leq)$.

- $x \epsilon X$ is the **upper bound** of A *iff* $a \leq x$ for all $a \epsilon A$.
- $x \epsilon X$ is the **lower bound** of A *iff* $x \leq a$ for all $a \epsilon A$.
- $x \epsilon A$ is the **greatest element** of A *iff* $a \leq x$ for all $a \epsilon A$.
- $x \epsilon A$ is the **least element** of A *iff* $x \leq a$ for all $a \epsilon A$.
- $x \epsilon X$ is the **greatest lower bound** of A *iff* A has a nonempty lower bounds set and x is the greatest element of this set.
- $x \epsilon X$ is the **least upper bound** of A *iff* A has a nonempty upper bounds set and x is the least element of this set.

Lattice is a poset $(X, \leq)$ where any pair of X has a supremum and an infimum. **Meet** (infimum, represented by $\wedge$ or $\cap$) and **join** (supremum, represented by $\vee$ or $\cup$) are lattice operations (Gallier, 2014).

Let assume $\{x, y\}$ is a set, $z = x \wedge y$, meet of x and y represents greatest lower bound of the set $\{x, y\}$. Therefore $z \leq x$ and $z \leq y$ must be satisfied.

Let assume $\{x, y\}$ is a set, $z = x \vee y$, join of x and y represents least upper bound of the set $\{x, y\}$. Therefore $x \leq z$ and $y \leq z$ must be satisfied.

According to definitions mentioned below, properties of lattice, written below, can be easily derived (Border, 2011):

- $x \wedge y \leq x \leq x \vee y$
- If $y \leq x$, then $x = x \vee y$ and $y = x \wedge y$
- $x = x \wedge x = x \vee x$
- $x \wedge y = y \wedge x$ and $x \vee y = y \vee x$
- $x \vee y \leq z$ *iff* $x \leq z$ and $y \leq z$ and $z \leq x \wedge y$ *iff* $z \leq x$ and $z \leq y$

Hasse diagram is a representation method of ordered sets by showing hierarchy in the diagram (Brüggemann and Patil, 2011). Assume that $x = \{a\}$, $y = \{a, b\}$, $z = \{a, c\}$ and $t = \{a, b, c\}$ are subset of X, where X is a poset $(X, \leq)$ and X is power set of $\{a, b, c\}$. If there is a relation between x and y such as $x \leq y$, where $\leq$ is a subset relation, then x is drawn below y in the vertical axis and the same for the other relations between pairs ($x \leq z$, $y \leq t$ and $z \leq t$) (Brüggemann and Patil, 2011). Hasse Diagram of the example is shown in the Figure 2.1.

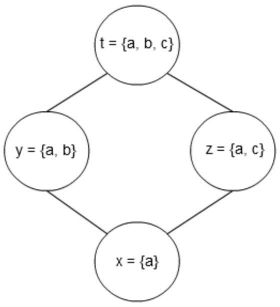


Figure 2.1. Hasse Diagram example represented with set name and elements of the sets.

2.2.2. Formal Concept Analysis

Formal Concept Analysis (FCA) is proposed by Will Rudolf and defined as restructuring lattice theory to conceptualize the lattice structure by adding the properties of objects to extend theory for the content of the data (Wille, 2009). In the original paper Wille (2009), FCA (or restructuring lattice theory) is described as "an attempt to reinvigorate connections with our general culture by interpreting the theory as concretely as possible, and in this way to promote better communication between lattice theorists and potential users of lattice theory." (Ganter et al., 2005).

Formal Concept Analysis in the original paper of Wille (2009), starts with a **formal context** definition which is a triple (G, M, I). G and M are sets and I is a binary relation represented as $I \subseteq G \times M$ between G and M (Yevtushenko, 2004). Elements of G are referred to as objects and elements of M are referred to as attributes (Yevtushenko, 2004).A formal context is actually an object-attribute relation where object consists of attributes. You can see formal context example in the following table (example taken from paper Yevtushenko (2004)): In Table 2.1, each row (mulled wine, coke, tee, coffee, mineral water) represents different objects and each column (soft, strong warm, sparkling, with caffeine) represents different attributes. For example, attributes of mulled wine is strong and warm, these are the properties which define/describe the mulled wine. Since formal concept is a binary relation between objects and attributes we can show an example formal context for tee. Formal context is a triple (G, M, I). G represents objects which are mulled wine, tee, coffee, $\ldots M$ represent attributes which are soft, strong, $\ldots I$ is the binary relation between G and M.

Table 2.1. Formal Context Example

	Soft	Strong	Warm	Sparkling	With caffeine
Mulled wine		X	X		
Coke	X			X	X
Tee	X		X		X
Coffee	X		X		X
Mineral Water	X				

Derivation Operator ('):

This operator represents a correspondence relation between a set of objects A and the set of attributes that is common in every object of A or vice versa; a set of attributes of B and all objects from G that contains all attributes of B (Yevtushenko, 2004). Derivation operator definition for object set:

Let $A \subseteq G$. Then $A' = \{m \epsilon M \mid gIm \quad \forall g \epsilon A\}$ (Yevtushenko, 2004)

Example of derivation operator on coffee object:

$\{Coffee\}' = \{soft,\ warm,\ with\ caffeine\}$ (attributes of coffee)

Derivation operator definition for attribute set:

Let $B \subseteq M$. Then $B' = \{g \epsilon G \mid gIm \quad \forall m \epsilon B\}$ (Yevtushenko, 2004)

Example of derivation operator on warm attribute:

$\{Warm\}' = \{mulled\ wine,\ tee,\ coffee\}$ (objects can be represented by warm attribute)

Closure Operator (''):

Closure of a set of attributes B, where $B \subseteq M$, consists of all attributes that are common in the objects that contain all attributes of B and vice versa; closure of a set of objects A, where $A \subseteq G$, consists of all objects that can be represented by common attributes that exist in all objects of A (Yevtushenko, 2004)

Example of closure operator on coffee object:

$Coffee' = \{soft,\ warm,\ with\ caffeine\}$ and $\{soft,\ warm,\ with\ caffeine\}' = \{tee,\ coffee\}$

$Coffee'' = \{tee, coffee\}$

Example of closure operator on warm attribute:

$Warm' = \{mulled\ wine,\ tee,\ coffee\}$ and $\{mulled\ wine,\ tee,\ coffee\}' = \{warm\}$

Therefore $warm'' = \{warm\}$

A set x is called **closed set** *iff* closure of it equals to itself ($x'' = x$).

A formal concept can be defined as a pair (A, B) where $A \subseteq G$, $B \subseteq M$ and $A' = B$, $B' = A$. A is called **extent** and B is called **intent** of the concept (Yevtushenko, 2004). Here are two examples of formal concept, where extents represent objects (coke, tee, . . .) and intents represent attributes (soft, sparkling, . . .).

1. $(\{Coke\},\ \{Soft,\ Sparkling,\ With\ caffeine\})$
2. $(\{Tee,\ Coffee\},\ \{Soft,\ Warm,\ With\ caffeine\})$

A formal concept is a pair of extents and intents but not as random combination of them. Each concept represents combination of extents and intents. In a concept, extents are combined based on common intents and vice versa, intents are combined based on being common in the same extents. Therefore, if a concept is represented as a pair (A, B), where A is extent set and B is intent set of the concept, intent set of a concept is derivation of extent set of the same concept : $B = A'$ and vice versa, extent set of a concept is derivation of intent set of the same concept : $A = B'$. Additionally, if closure of extent set of a concept is equal to itself, $A'' = A$, extent set is called **closed set**.

There is a **subconcept-superconcept** relation between concepts of a context which is also a representation of partial order: $(A1, B1) \leq (A2, B2) \iff A1 \subseteq A2$ (at the same time $B2 \subseteq B1$) and $\leq$ is called hierarchical order in FCA (Yevtushenko, 2004).

For example:

$(\{Coke,\ Tee\},\ \{Soft,\ With\ caffeine\}) \leq (\{Coke,\ Tee,\ Mineral\ water\},\ \{soft\})$

Concept lattice is the set of all objects of the context, ordered by the subconcept-superconcept relation and represented by $\underline{\beta}(G, M, I)$.

CHAPTER 3

RELATED WORK

Although there are various work on the news analysis in general, “coherent news chain construction” is relatively a new area. The aim of coherent news chain creation is to find news chains consists of time-ordered news chains which have coherence through the whole story. There are different studies which try to create relation between news articles and to discriminate news articles according to their topics, which are classified under Topic Detection and Tracking. However, the aim of news chain creation focuses on coherence through all news articles that exist in the chain, not just finding related documents.

Topic Detection and Tracking (TDT) is an initiative of which has the aim “to investigate the state of the art in finding and following new events in a stream of broadcast news stories” (Allan et al., 1998). Three tasks are defined in the scope of TDT which are (Allan et al., 1998):

- The Segmentation Task: Aims to discriminate stories in the corpus according to related topic.
- The Detection Task: Aims to detect events and if it already exists or a new event.
- The Tracking Task: Aims to relate an incoming event with already existing event.

The Segmentation Task and The Tracking Task can be used as a helper in coherent news chain construction, but they are not focusing on creation chain as a whole story. Indeed, there are differences in terminology. In TDT approaches, each document is accepted as a story which may include more than one event. However, in news chain construction, story is accepted as a chain of news which creates a semantic flow of subject.

After relation between TDT and coherent news chain construction is explained above, we can focus on the approaches suggested for news chain construction.

For news story and chain construction, different approaches are suggested. We can classify these approaches under three titles: Probabilistic approaches, linear programming approaches and graph-based approaches. Our method can be classified as a graph-based approach.

Methods suggested by Ahmed et al. (2011) and Gillenwater et al. (2012) are probabilistic approaches which try to create a solution for news chain construction problem by using probabilistic distributions.

In the study "Unified Analysis of Streaming News", Ahmed et al. (2011) suggest a unified methodology to apply grouping news articles into storylines by extracting topic of the articles based on a probabilistic method. Suggested method is a hybrid method that covers clustering algorithms and topic models. In the article, topic is considered as a loose relation for a document in a long-time scale while story lines are assumed to focus on event or actors in short-time interval. In the study Recurrent Chinese Restaurant Process is used for based clustering method is used extended by topic information, that is determined by using Latent Dirichlet Allocation Model, to learn new information from news stream. Then, named entities are used to be able to distinguish storylines, which are not general as topic, to from each other. And finally with a non-parametric probabilistic approach, which is Recurrent Chinese Restaurant Process Model, a probability distribution is created by using timestamps.

In the study "Discovering Diverse and Salient Threads in Document Collections", Gillenwater et al. (2012) suggest a probabilistic model to extract diverse set of paths from dataset of news articles or academic papers. To achieve this task, collection is transformed into a directed graph where nodes represent documents and edges represent relations between documents. Weights are assigned by a function in which weights of nodes are scores for document importance and weight of edges are relative strength of relation between articles. After a directed graph representation and weight assignment to nodes and edges are constructed, a model which is a combination of *SDPPS* (structured determinantal pont processes) and *k-DPP* (fixed size determinantal point processess) is used as a probabilistic model to find diversity in document collection.

Linear programming solutions generally are based on pairwise coherence for news chain creation. Pairwise coherence is calculated between news articles to create consecutive news articles of news chain. For the task, advantages of linear programming are used.

In the paper "Connecting the Dots Between News Articles", Shahaf and Guestrin (2010) suggest a methodology for automatically connecting the dots by using a linear programming solution for coherence problem. The aim of the study is to construct coherent news chains defined by the two endpoint news articles that represent start and end of the chain. To accomplish the task, a linear programming-based coherence measurement is used to assign coherence scores to each news article sequence. One of the key points of the project is the usage of word frequencies to determine the active (pairwise occurrence evaluation) and init (difference between consecutive pairwise occurrence evaluations) values for words appearing in the chain articles. Rather than the discrete existence of the keyword, keywords that appear and disappear just in a few news articles, keywords that exist through the chain consistently have a main effect on the scoring of coherence. Therefore, the main idea of the paper is to preserve global coherence by the use of activation and initialization patterns of the words with the objective of maximizing the weakest pairwise (local) article coherence. To determine influence of words in the chain, a bipartite graph is used.

In the paper “Trains of Thought: Generating Information Maps”, Shahaf et al. (2012b) suggest a methodology to create metro maps that are structured summaries of information. According to the paper, search engines are effective enough to retrieve documents related with search keywords, but they are far from creating a comprehensive big picture of the topic. In the study, candidate news chains are created based on the coherence measurement as declared in the paper “Connecting the Dots Between News Articles”, mentioned in previous paragraph, which is a linear programming solution, and then extending it with coverage and connectivity parts, thus news chains are produced that share the same topic but with its different aspects.

In the paper “Metro Maps of Science”, Shahaf et al. (2012a) modify their study “Trains of Thought: Generating Information Maps”, mentioned in previous paragraph, for the creation of metro maps of scientific articles. Methodology is modified by interrogating coherence in scientific papers by extending the content of data with the inclusion of citations.

There are also graph-based approaches for news chain creation. The intuition of these approaches is generally based on the relation between news articles. Relations between news articles are structured as a graph and by pruning or reducing the paths of the graph, best candidates are determined.

In the study “Generating Pictorial Storylines Via Minimum-Weight Connected Dominating Set Approximation in Multi-View Graphs” (Wang et al., 2012), a method to construct pictorial and structural summarization of a topic is suggested. Dataset of the method consists of images described by short text. The key idea of the method is constructing a graph accepting images as vertices, undirected edges as object similarities and directed edges as pairwise object temporal relations. After creation of the graph, a minimum-weight dominating set algorithm is applied to find most representative objects of the topic and then Steiner tree approximation algorithm is applied to construct a storyline by using representative objects. In the study, a weighted graph is used to create a pictorial storyline which differs from our study in usage of graph representation. In our approach, lattice structure is used which creates a hierarchy between nodes instead of a regular graph representation.

Lattice is a specific graph structure and used in the news chain/story construction. In concept lattice structure, general intention is conceptualizing the hierarchy between news articles. There are suggested methods which uses concept lattices as a solution.

In the study "A combined approach of formal concept analysis and text mining for concept based document clustering", Myat and Hla (2005) suggest a method for concept based document clustering. In the scope of the project, a concept lattice is produced as a result based on relation between documents and terms. Documents are accepted as objects and terms are accepted as attributes. For reduction of high dimensionality of terms, they use tf-idf (term frequency-inverted document frequency) weighting to get more relevant terms. They try to find association between terms which are filtered by using a threshold on the lattice structure. By using some threshold techniques, concepts which are not belong to a conceptual cluster are eliminated. Since, this method uses concept lattice, it is a related work. They construct lattice by using documents-terms relations. But the aim of the study is clustering documents where our study differs from it.

In the paper "Reasoning about Sets using Redescription Mining" (Zaki and Ramakrishnan, 2005), a redescription mining approach is proposed. Redescription mining is an approach tries to find different expressions of the same subset which have duality since they are representing the same information. In the study, Zaki et al. suggest a methodology that constructs lattice by using closed itemsets and then minimal generator that constructs closed sets and final non-redundant redescriptions of minimal generator are generated. This method aims to generate non-redundant description of big data to create more understandable pattern of the data. For lattice construction, CHARM and CHARM-L methods are used.

In the paper "CHARM: An Efficient Algorithm for closed Association Rule Mining", Zaki and Hsiao (2002) suggest an algorithm for association mining rule that finds non-redundant association rules without mining all frequent itemsets. According to study, examining all frequent itemsets is not obligatory to find all non-redundant association rules. All non-redundant association rules can be explored by mining just closed itemsets (closure operation is mentioned in the Chapter 2). Usage of closed itemsets will save computation time by ignoring examination of frequent itemsets that do not satisfy closure operator. However, identifying closed itemsets is still an expensive computation. To reduce computation time of identification of closed itemsets, exploration of the itemset space and transaction space is applied at the same process time/level.

In the paper "Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure" (Zaki and Hsiao, 2005), CHARM-L algorithm is presented as a lattice construction approach. This approach is based on efficiency of CHARM algorithm mentioned above. Mining just closed itemsets gives computation power to CHARM. In addition, CHARM-L differs from other lattice construction methods by creating lattice during the closed itemset mining. It generates lattice structure during generation of closed itemsets. When a closed itemset is determined, possible closed subsets of the itemset is determined and computational efficiency is increased.

The work "Storytelling in Entity Networks to Support Intelligence Analysts" (Hossain et al., 2012) focuses on the creation of chains between entities. The aim is to create links between entities by using a graph construction and search methods such as Cover Tree Nearest Neighbor, Nearest Neighbors Approximation, and k-Clique Near Neighbor are used. Entities that are close to the initial entity which is desired to be the start of the story are extracted as candidates. Finally, candidates are filtered according two different modes called normal and mixed modes respectively. Created graph in the study consists of cliques which represent the neighbors of the sequential news. By using concept lattice as an optimization structure, stories are constructed rapidly in massive data. In the study, graph consists of cliques and lattice usage is not explained in details. In our method, lattice is not used as an optimization method; in fact lattice is used to construct relation between news articles.

In the study "Helping Intelligence Analysts Make Connections", Hossain et al. (2011) created a visualization system based on a storytelling algorithm. This system tries to help intelligence analysts with an automated system which creates connection between documents/news articles to ease the analysis on large datasets. In addition to this, system also allows user to modify news chains according to his/her wishes. In the paper, there are two predetermined documents as start and end articles and problem is finding related documents which create a coherent chain between them.

In the method, beginning from the start article, at each step a set of neighbour articles is determined to construct a graph. The neighbourhood information is extracted from a concept lattice. The concept holding the start article is examined to get the list of other articles that are related to it through the use of distance measurement (Soergel distance) and the longest term set. The list of articles that should be connected to the previous node is determined according to clique size. Thus, distance measurement between articles and the clique size (the number of neighbours to select at each step) are the two constraints of the method.

To summarize, using distance threshold and clique size between consecutive documents in the graph, possible successors are found. Finally, using A* search algorithm with Soergel distance, optimal path from the graph, including the documents in clique structure, is determined as the best chain between predetermined documents. This study can be called as the improved version of the study "Storytelling in Entity Networks to Support Intelligence Analysts" (Hossain et al., 2012). A graph is constructed which consists of cliques, as in their previous study (Hossain et al., 2012), but lattice is used to find possible neighbors of the start article of the story. Difference between our method and this method is the usage of lattice. In the method, lattice is used as a helper to narrow down the set of possible news articles which create a clique around sequential news articles and aim of the project is to find the best news chain between predetermined start and end articles. Our method, on the other hand, aims to find all possible coherent news chains, not just the one between two predetermined articles, in whole dataset and concept lattice is the backbone of our study not just an optimization method.

CHAPTER 4

MATERIAL AND METHOD

Intuition of the method is based on the exploration of partial order of news articles that share common words related with a topic. Lattice structure, as a specific partial order representation, is used to explore hidden partial order relation between news articles. Inverted index is used to construct a concept lattice. Secondary keys of document space/set, which are terms, are used as extents and posting list elements, which are document names, are used as the intents of the concepts.

Contribution of the method is the proposed use of inverted index for the creation of partial order of news articles. There is no similar study which uses inverted index structure of document set to create concept lattice based on terms of articles. This approach does not require predetermination of fixed start and end articles of the news chain. There is already time order in the corpora used in the study, therefore using terms that chain will be constructed upon aims to create a coherent news chain without a predefined start and end articles. After a news chain is created with method suggested by us, start and end article of a chain is determined according to time order of articles. Therefore, start and end article of chains is not known since intent set (document set) of the concept is not known before it is constructed. Aim of the “connecting the dots” approach is to find news articles which creates coherent chains between predetermined two articles. Since, suggested method by us is aimed to create all possible news chains, not just related ones with predetermined two articles, which exists in a given dataset, our method cannot be classified as one of the “connecting the dots” approaches.

In this section, a detailed description of the methodology will be given. General steps of the method will be described step by step and at the end of the section, steps are practically shown on an example dataset.

4.1. Dataset

Two commonly used news article datasets are selected for the evaluation of the method: The New York Times Annotated Corpus (Sandhauts, 2008) and The Reuters Corpus Volume 1 (Rose et al., 2002). The New York Times Annotated Corpus is prepared by using the archive between January 1, 1987 and January 1, 2007. The Reuters news archive covers dates between August 20, 1996 and August 19, 1997. Both corpora represent news articles and their metadata (additional information like topic of the article) by using their own ontology created in Extensible Markup Language (XML) structure. The New York Annotated Corpus contains approximately 1,800,000 articles and The Reuters Corpus Volume 1 contains approximately 800,000 articles.

Since both corpora cover different time intervals, subsets of news articles are selected according to publication time to create the same time interval for the comparison of created chains. Therefore, articles between August 20, 1996 and August 19, 1997 are selected from both corpora to work on the same time interval for the datasets.

4.2. Topic Selection and Dataset Filtering

Both corpora include big amounts of news articles on different topics. Since the aim of the study is to produce coherent news chains, articles are filtered according to a selected topic because it is more likely that a news chain will be more coherent if its articles share the same topic like “economics”. The other reason of filtering news articles according to a selected topic is to reduce computational time of the method for the case study since lattice construction is computationally expensive. Therefore, as an initial step, “politics” is determined as a general topic and articles that are assigned to topic “politics” are extracted from the corpora in the time interval August 20, 1996 and August 19, 1997. In this report this filtered dataset will be called subsets “politics”. Since there is no unique topic for politics in both datasets, for example there are topics in The Reuters Corpus Volume 1 like “domestic politics”, “internal politics” and “current news - politics” and all of them are related with politics, a manual topic selection is applied. Selected topics that are related with politics both for The New York Times Annotated Corpus and The Reuters Corpus Volume 1 are listed in APPENDIX A.

After filtering datasets according to politics related topics, 6676 news articles remained in The New York Times Annotated Corpus subset and 183254 news articles in The Reuters Corpus Volume 1 subset.

4.3. Document Parser

Both corpora, used in the study, have their own structure for the representation of news articles and they also provide individual parser codes written in Java. We added an abstract layer to the already provided parser codes, in other words we abstracted document structure to represent common parts of the articles from both datasets. Title, body, publication date and categories are common parts used in the study.

In addition to this, some of the already existing libraries are used for some basic natural language processing methods like stemming and lattice construction.

4.4. Stop Words Removal

Relation between words and documents is the main focus in the study. Each word is accepted as a single and independent element of the document, therefore no phrases are used. Stop words generally includes specific meaning in the phrases and since no phrases are used in the study and each single word is accepted as independent unit of documents, stop word removal is used to discard words which commonly does not imply specific meaning for documents.

Union of two stop words lists, Onix Text Retrieval Toolkit Stop Word List 1 (`http://www.lextek.com/manuals/onix/stopwords1.html`) and Onix Text Retrieval Toolkit Stop Word List (`http://www.lextek.com/manuals/onix/stopwords2.html`), is used for stop words removal.

4.5. Stemming

Stemming algorithms are generally heuristic approaches that try to find the stem of the word (Manning et al., 2008). Since they are heuristics approaches, they do not guarantee to find morphologically correct stem of the word. Therefore, the error rate of stemming algorithms must be accepted in use of them.

Porter Stemmer (Porter, 1997), which is used in the study as the stemming method, is an approach for stemming that removes suffixes from words to reduce total number of words and complexity of the data to improve efficiency of information retrieval systems. It consists of 5 steps as suffix removal that can be found in the paper (Porter, 1997).

Processed words, words on which natural language methods are applied, are called terms in the study.

4.6. Inverted Index Creation

Stemming is applied to words that remained after stop words are removed and then inverted index is created. Structure of inverted index includes term, total frequency of term in dataset and document names that term occurs in.

Example :

Clinton 7888 872452;872473;872478;872515;872575;872591;872595;...

Here Clinton is term itself, 7888 is the total frequency of term “Clinton” in the whole dataset and 872452;872473;872478;872515;872575;872591;872595;... are documents that term “Clinton” occurs in.

Inverted index structure is stored in a file system that contains separate files based on the first letter of the term. It means, terms that start with letter ’a’ are stored in a file named “a”. For example Clinton is stored in file “c” and term “galileo” is stored in file named “g”. This approach is preferred to fasten the access of inverted index of a term since it is not always possible to store inverted indices of all terms in physical memory.

4.7. Term Frequency- Inverse Document Frequency Calculation

Term Frequency (TF) is the total occurrence of the term in all documents of dataset. This value is simultaneously calculated during the inverted index creation step and is stored in the inverted index file structure.

Inverse document frequency (IDF) is calculated by using a ratio between document number of dataset and document number that term occurs in. IDF of a term is calculated after inverted index is constructed, because posting list of the term, document names of a term occurs in, is used to calculate.

$$idf_t = \log \frac{N}{df_t} \tag{4.1}$$

where N is the number of documents in dataset and df_t is the number of documents that the term exists in.

4.8. Term Selection

Term frequency and inverse document frequency are calculated for the ranking of the terms in the datasets. For the first evaluation of the study instead of focusing on tuning of term selection, first feedback of the method has been the focus. Therefore according to the ranking of terms by both TF and IDF values, top terms are selected to create news chains based on them. Since number of news chains will be large because of the datasets' sizes, maximum number of terms are tried to be selected to decrease number of produced news chains to make manual check of coherence easier for the case study. Maximum number of terms, to create news chains based on themes around these terms, with an acceptable time consumption for lattice construction is determined as 10. So approach is applied on top 10 terms.

4.9. Concept Lattice Construction

After terms are selected, concept lattice is constructed by using CHARM-L method (Zaki and Hsiao, 2005) which is explained in Section 3. For this task, a java library entitled as Galicia (Valtchev et al., 2003) is used. To use CHARM-L method of Galicia library, it is required to supply a binary relation table (could be plain text file or Extensible Markup Language (XML) formatted file) as input. As a result XML formatted output is produced which includes concepts, extent and intent sets of the concepts and superconcept relations between concepts.

4.10. Example of Steps

Here an example dataset will be defined and process steps will be applied on it to show the workflow of the method clearly.

Let's say that we have a document set shown in Table 4.1. There are 7 documents listed in the first column and terms contained by the matching document are put in the second column. For example document with id d1 contains words "School, School, Education, Student, Effect, And, Or, Before".

Table 4.1. An example dataset that contains documents and words contained by them

Document ID	Content of Document
d1	School, School, Education, Student, Drug, And, Or, Before
d2	School, Effect, Education, Student, Drug, Precaution, Too, In
d3	Effect, Effect, Student, Drug, Precaution, Sense, Also
d4	School, Education, Drug, Precaution, An, Also
d5	School, Effect, Education, Student, Drug, Be, Or
d6	School, Education, Student, Drug, May, More
d7	Effect, Student, Drug, Feel, About, An

Let's assume we have a stop word list shown in Table 4.2. All of the words listed below also exist in the Onix Text Retrieval Toolkit Stop Word List 1.

Table 4.2. An example of stop words list

Stop Word List
And
Also
An
About
Be
Before
In
May
More
Or
Too

In the stop word removal step, each word in the document is compared with words in the stop word list and if there is a match, that word is removed from the document content. After example stop word list shown in Table 4.2 is processed on each example document shown in Table 4.1, document content remains as shown in Table 4.3. As an example, in the first document with id d1, words “And, Or, Before” are removed.

Table 4.3. Example document set after stop words are removed

Document ID	Content of Document
d1	School, School, Education, Student, Drug
d2	School, Effect, Education, Student, Drug, Precaution
d3	Effect, Effect, Student, Drug, Precaution, Sense
d4	School, Education, Drug, Precaution
d5	School, Effect, Education, Student, Drug
d6	School, Education, Student, Drug
d7	Effect, Student, Drug, Feel

After stop word removal step, Porter Stemmer algorithm is applied on each word of the documents. Stemming algorithm modified word “Education” to stem “Educ”, word “Precaution” to stem “Precaut” and word “Sense” to stem “Sens”. New document dataset after stemming process is shown in Table 4.4.

Table 4.4. Example document set after stop words are removed

Document ID	Content of Document
d1	School, School, Educ, Student, Drug
d2	School, Effect, Educ, Student, Drug, Precaut
d3	Effect, Effect, Student, Drug, Precaut, Sens
d4	School, Educ, Drug, Precaut
d5	School, Effect, Educ, Student, Drug
d6	School, Educ, Student, Drug
d7	Effect, Student, Drug, Feel

Then, inverted index is constructed and at the same time term frequency is calculated for each term. Then using posting list, inverse document frequency is calculated for each term. Term, TF and IDF values are shown in Table 4.5.

Table 4.5. Inverted index of example dataset

Term	TF - IDF	Posting List
School	6 - 0.14612	d1, d2, d4, d5, d6
Effect	5 - 0.24303	d2, d3, d5, d7
Educ	5 - 0.14612	d1, d2, d4, d5, d6
Student	6 - 0.06694	d1, d2, d3, d5, d6, d7
Drug	7 - 0	d1, d2, d3, d4, d5, d6, d7
Precaut	3 - 0.36797	d2, d3, d4
Sens	1 - 0.84509	d3
Feel	1 - 0.84509	d7

If we sort terms according to the decreasing order of their TF, top 6 keywords will be “Effect”, Student, “Educ”, “School”, “Drug”, “Precaut”. These are selected as terms that news chain will be constructed on.

By using inverted index of these keywords, closed itemsets will be mined to find the concepts of the lattice. In the formal concept analysis, a concept consists of extents (combination of objects) and intents (combination of attributes) which are terms and documents that term occurs in. In the example, terms represent objects and documents represent attributes.

A closed itemset is the subset of objects which is equal to derivation of common attributes of itself as mentioned in Section 2.2.2. For example:

Let $x = \{Drug, Precaut\}$ be a subset of X which is all object set that is equal to $\{School, Effect, Educ, Student, Drug, Precaut, Sens, Feel\}$

x is a closed set *iff* x'' (closure of x) is equal to x as mentioned in Section 2.2.2.

To see this, let y be derivation of x; that is:

$y = x' = \{d2, d3, d4\}$

Since $x'' = (x')'$ mentioned in Section 2.2.2, we have:

$x'' = (x')' = y' = \{Drug, Precaut\}$

Thus $x'' = x$ and x is a closed set.

All closed itemsets are explored by using top 6 terms and shown in Table 4.6.

Table 4.6. All closed itemsets of the top 6 terms

Concept name	Extents	Intents
Concept 1	Drug	d1, d2, d3, d4, d5, d6, d7
Concept 2	Drug, Precat	d2, d3, d4
Concept 3	Student, Drug	d1, d2, d3, d5, d6, d7
Concept 4	School, Educat, Drug	d1, d2, d4, d5, d6
Concept 5	Effect, Student, Drug	d2, d3, d5, d7
Concept 6	School, Educat, Student, Drug	d1, d2, d5, d6
Concept 7	Effect, Student, Drug, Precat	d2, d3
Concept 8	School, Effect, Educat, Student, Drug	d2, d5
Concept 9	School, Educat, Drug, Precat	d2, d4
Concept 10	School, Effect, Educat, Student, Drug, Precat	d2

After all concepts are explored in the dataset, concept lattice output is produced. This output is generally represented in a structured text format like XML or as a diagram like Hasse Diagram. Hasse Diagram of the concept lattice of the example dataset is shown in Figure 4.1. Concepts are listed in a hierarchical order in the diagram. For example, concept2, which is $Drug, Precat$, should be above the concept1, which is $drug$, in the vertical axis because concept2 is a super concept of concept1 since concept1 $\leq$ concept2.

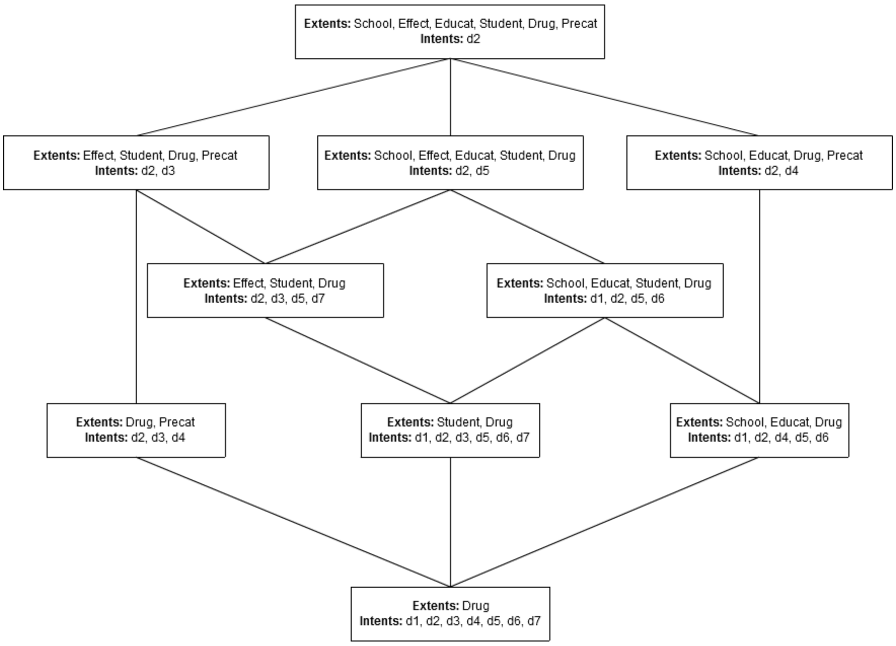

Extents: School, Effect, Educat, Student, Drug, Precat
Intents: d2
Extents: Effect, Student, Drug, Precat
Intents: d2, d3
Extents: School, Effect, Educat, Student, Drug
Intents: d2, d5
Extents: School, Educat, Drug, Precat
Intents: d2, d4
Extents: Effect, Student, Drug
Intents: d2, d3, d5, d7
Extents: School, Educat, Student, Drug
Intents: d1, d2, d5, d6
Extents: Drug, Precat
Intents: d2, d3, d4
Extents: Student, Drug
Intents: d1, d2, d3, d5, d6, d7
Extents: School, Educat, Drug
Intents: d1, d2, d4, d5, d6
Extents: Drug
Intents: d1, d2, d3, d4, d5, d6, d7

Figure 4.1. Hasse diagram of the example dataset's concept lattice.

CHAPTER 5

RESULTS

We selected lattice construction by using the most frequent 10 keywords (will be called top 10 keywords), mentioned in Section 4.8, as the case study. These keywords are belong to politics subdataset, mentioned in Section 4.2, of The New York Times Annotated Corpus. Using politics subdataset, a concept lattice is constructed by assuming the top 10 keywords as objects of the context. As a result of the process, a lattice is produced with 1024 concepts in which each concept is a unique term combination and a closed set . It is predictable when we consider all possible combinations of 10 keywords which is 1024 with the addition of empty object set. Since it is not visually effective to show it, hasse diagram of concept lattice is not given in this report.

Top 10 keywords according to term frequency in politics subdataset are listed in the Table 5.1. Also, top 100 keywords according to the term frequency in politics subdataset can be seen in APPENDIX B to gain an insight on frequent keywords related with politics of subdataset.

Table 5.1. Top 10 keywords of politics subdataset of New York Times Annotated Corpus

Order	Keyword	Frequency
1	Govern	11266
2	Presid	10278
3	Offici	9069
4	American	8838
5	Polit	8707
6	Unit	8623
7	Clinton	7888
8	Peopl	7430
9	Nation	7423
10	Parti	7101

Rather than providing the visualization of the concept lattice, distribution statistics of extents of the concepts are used to analyze lattice. For the creation of the extent distribution of the lattice, relative extent size (Yevtushenko, 2004) and relative frequency (Yevtushenko, 2004) properties are used. According to study Yevtushenko (2004), relative extent size is defined as:

$$relative\_extent\_size = \frac{|A|}{|G|} \tag{5.1}$$

where A represents extent set of the concept, $|A|$ represents size of the extent set of the concept and G represents set of all extents of the context, $|G|$ represents number of all the extents of the context (Yevtushenko, 2004). Relative extent size is between $[0, 1]$ range since it is a ratio between a subset of extent set and size of extent set of the context. Relative frequency is formulated as:

$$relative\_frequency = \frac{|(A, B)||A| = s|}{|\underline{\beta}(G, M, I)|} \tag{5.2}$$

which is the ratio between relative extent size frequency and total concept number. In the formula,$|(A, B)||A| = s|$ represents occurence number of relative extent size s and $|\underline{\beta}(G, M, I)|$ represents the size of concept lattice, G is object set, M is attribute set and I is relation between objects and attributes of concept lattice .

Relative frequency vs. relative extent size of the lattice produced by top 10 keywords of politics subdataset can be seen in Figure 5.1. It has a normal distribution because the produced lattice includes every combination of objects, which is 1024 concepts for 10 keywords. For example, the combination that includes all 10 keywords, will exist once in the lattice and its relative extent size is $\frac{10}{10} = 1$, the dividend number is extent number of the concept and the divisor is total object number in the context. And relative extent frequency is $\frac{1}{1024} = 0.00098$, 1 is the relative extent frequency of concept that contains all extents and 1024 is the total concept number.

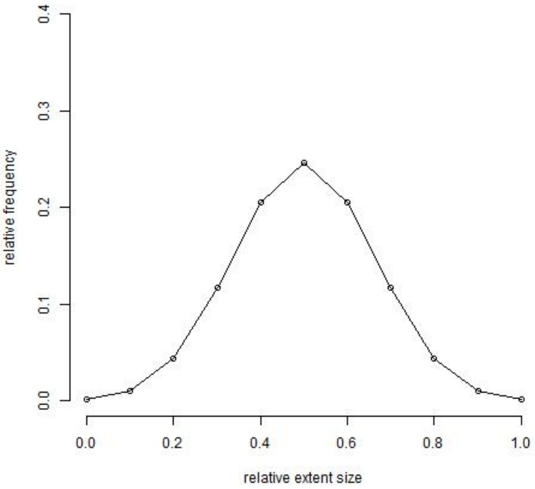


Figure 5.1. The relative frequency vs. relative extent size of the lattice produced by using top 10 keywords of politics subdataset of The New York Times Annotated Corpus.

Each concept consists of extents and intents and the relation between them is defined by the closure operator. Depending on the closure relation between intent and extent, intents are accepted as descriptors of extents. Within the scope of this study; intents, which are documents, are the descriptors of extents, which are terms. Considering this, the intent set of a concept is determined as the candidates that contain news chains related with extents.

The produced concept lattice is too large to analyze thus there is a need for reduction. In order to see whether we can get promising containers of news chains, one path of the concept lattice is examined in the study, starting from the concept that includes all objects and ends at the concept with the empty extent set. Selected path of the concept lattice can be seen in Figure 5.2. Concept 1 includes all objects as extent set. By removing one object from the extent set of concept 1, the next concept is found. For example, difference between concept 1 and concept 1024 is the object called “american”. Since one object is removed from concept, the new concept will be in the lower level in the hierarchy. At the end of the path, concept with empty extent set remains. The reason to select this path is arbitrary, but there is a pattern in this path which is removing most frequent keyword while traversing path from concept1 which includes all terms as extent set through the concept4 which has empty extent set. But, as a future work, a traversal method is planned to constructed to evaluate all or most of the paths.

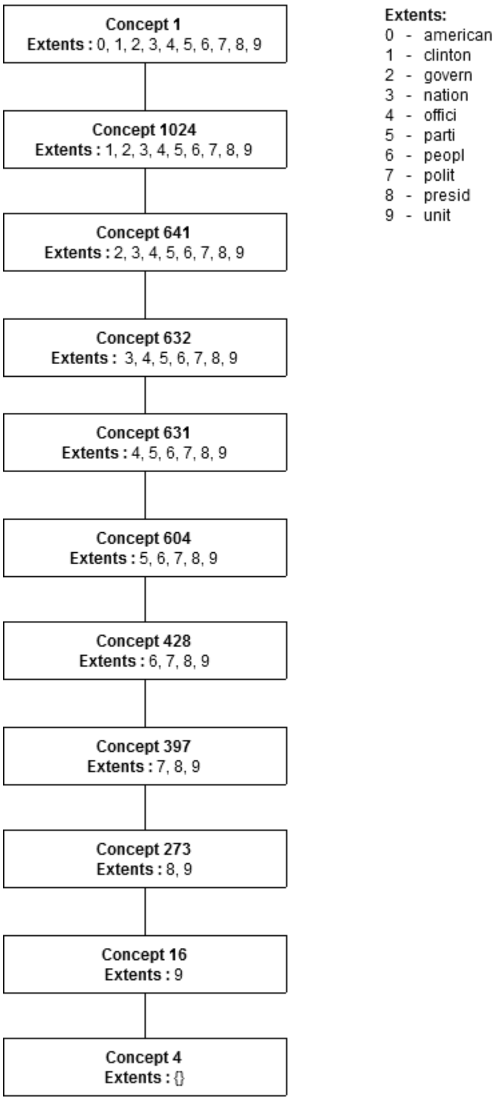


Figure 5.2. The selected path of the concept lattice and the extent mappings for the including concepts.

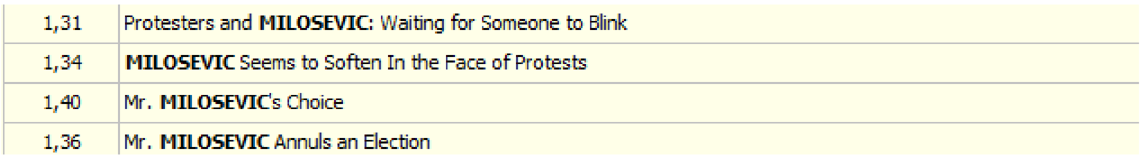
Intent set of each concept can be accepted as news chain candidates, but the path between concepts, which is intent differences, shows us theme changes through the path. Therefore, intent set difference between sequential concepts are accepted as news chain candidates. The sizes of intent differences between sequential concepts are listed in Table 5.2.

Intent differences between levels can also indicate the role of the specific extent (word) in unifying the set of those intents (documents).

Table 5.2. Number of intent differences between sequential concepts

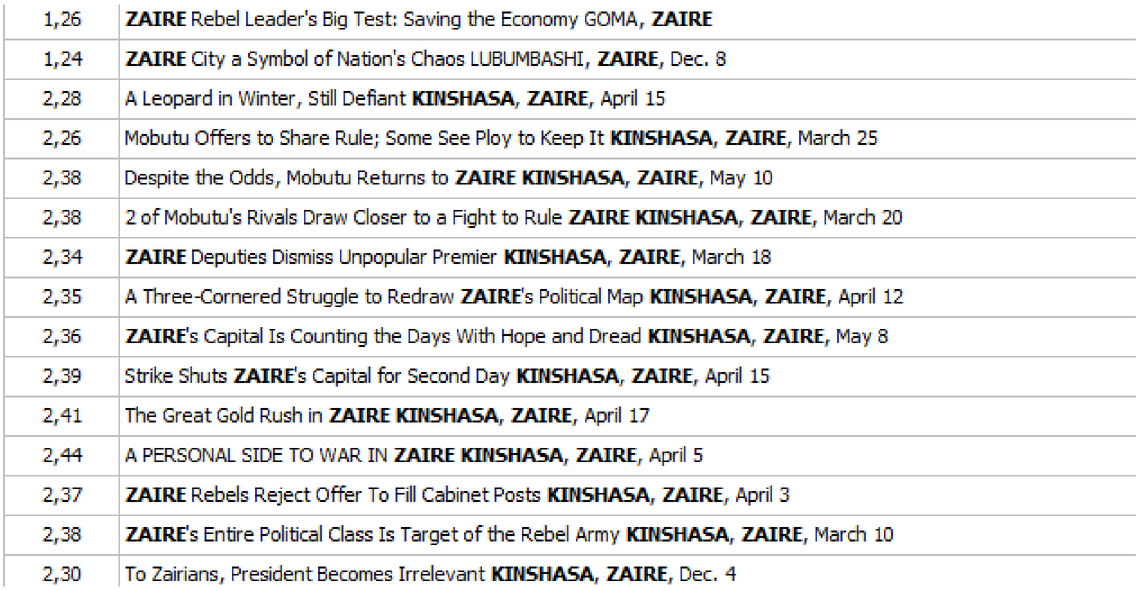
Concept pairs	Number of intent difference
Concept1 - Concept1024	15
Concept1024 - Concept641	110
Concept641 - Concept632	32
Concept632 - Concept631	46
Concept631 - Concept604	139
Concept604 - Concept428	276
Concept428 - Concept397	387
Concept397 - Concept273	608
Concept273 - Concept16	961
Concept16 - Concept4	2427

After selecting an example path from the lattice and determining the intent differences, sentence-based clustering is applied on document/news titles of the intent (document) differences between concepts. Each cluster contains news titles which are similar to each others which means similarly titled news are grouped as a cluster. Objects/elements of clusters are titles of news articles; therefore, similarity/distance between news titles determines clusters. By this, candidate news chains are discriminated according to their titles. The reason to apply clustering on news articles' titles is to discriminate news chains, because there may be more than one news chains which differs from each other based on theme. For example, there can be two news chains related with ”drug”, one is about ”drug” and its effect on youth culture and the other may be related with ”drug” and governmental precaution for drug . Since the size of intent differences is small and the transitions between concept 1 - concept 1024, concept 641 - concept 632 does not contain similar titles to create a cluster, no clusters could be produced based on these transitions. Selected clusters of other transitions are listed as examples below, not all clusters are shown here, just a subset is listed (Figures 5.3 - 5.16).



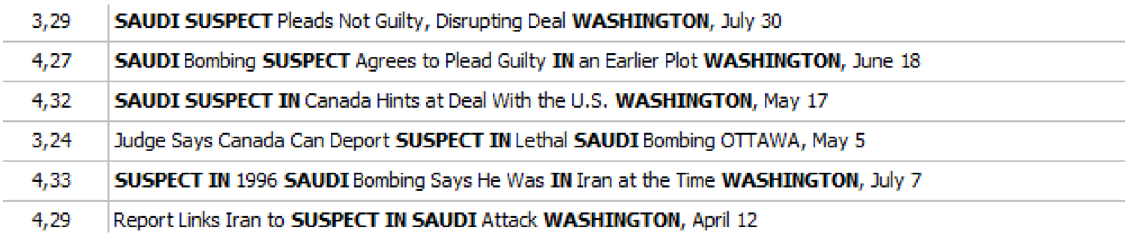
1,31	Protesters and **MILOSEVIC**: Waiting for Someone to Blink
1,34	**MILOSEVIC** Seems to Soften In the Face of Protests
1,40	Mr. **MILOSEVIC**'s Choice
1,36	Mr. **MILOSEVIC** Annuls an Election

Figure 5.3. First of two example clusters from transition between concept16 - concept4.



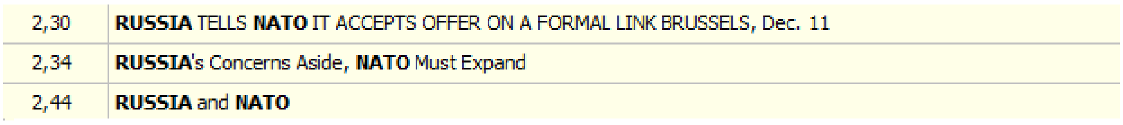
1,26	**ZAIRE** Rebel Leader's Big Test: Saving the Economy GOMA, **ZAIRE**
1,24	**ZAIRE** City a Symbol of Nation's Chaos LUBUMBASHI, **ZAIRE**, Dec. 8
2,28	A Leopard in Winter, Still Defiant **KINSHASA**, **ZAIRE**, April 15
2,26	Mobutu Offers to Share Rule; Some See Ploy to Keep It **KINSHASA**, **ZAIRE**, March 25
2,38	Despite the Odds, Mobutu Returns to **ZAIRE KINSHASA**, **ZAIRE**, May 10
2,38	2 of Mobutu's Rivals Draw Closer to a Fight to Rule **ZAIRE KINSHASA**, **ZAIRE**, March 20
2,34	**ZAIRE** Deputies Dismiss Unpopular Premier **KINSHASA**, **ZAIRE**, March 18
2,35	A Three-Cornered Struggle to Redraw **ZAIRE**'s Political Map **KINSHASA**, **ZAIRE**, April 12
2,36	**ZAIRE**'s Capital Is Counting the Days With Hope and Dread **KINSHASA**, **ZAIRE**, May 8
2,39	Strike Shuts **ZAIRE**'s Capital for Second Day **KINSHASA**, **ZAIRE**, April 15
2,41	The Great Gold Rush in **ZAIRE KINSHASA**, **ZAIRE**, April 17
2,44	A PERSONAL SIDE TO WAR IN **ZAIRE KINSHASA**, **ZAIRE**, April 5
2,37	**ZAIRE** Rebels Reject Offer To Fill Cabinet Posts **KINSHASA**, **ZAIRE**, April 3
2,38	**ZAIRE**'s Entire Political Class Is Target of the Rebel Army **KINSHASA**, **ZAIRE**, March 10
2,30	To Zairians, President Becomes Irrelevant **KINSHASA**, **ZAIRE**, Dec. 4

Figure 5.4. Second of two example clusters from transition between concept16 - concept4.



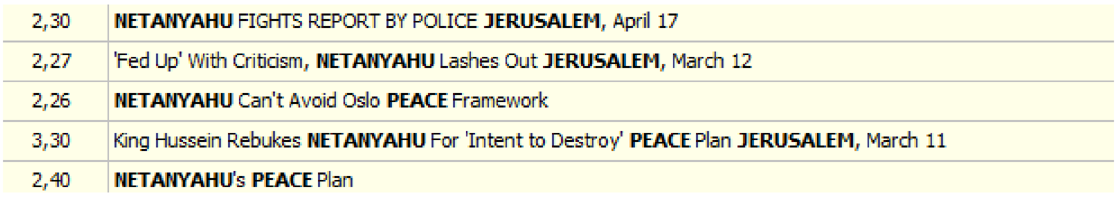
3,29	**SAUDI SUSPECT** Pleads Not Guilty, Disrupting Deal **WASHINGTON**, July 30
4,27	**SAUDI** Bombing **SUSPECT** Agrees to Plead Guilty **IN** an Earlier Plot **WASHINGTON**, June 18
4,32	**SAUDI SUSPECT IN** Canada Hints at Deal With the U.S. **WASHINGTON**, May 17
3,24	Judge Says Canada Can Deport **SUSPECT IN** Lethal **SAUDI** Bombing OTTAWA, May 5
4,33	**SUSPECT IN** 1996 **SAUDI** Bombing Says He Was **IN** Iran at the Time **WASHINGTON**, July 7
4,29	Report Links Iran to **SUSPECT IN SAUDI** Attack **WASHINGTON**, April 12

Figure 5.5. First of three example clusters from transition between concept273 - concept16.



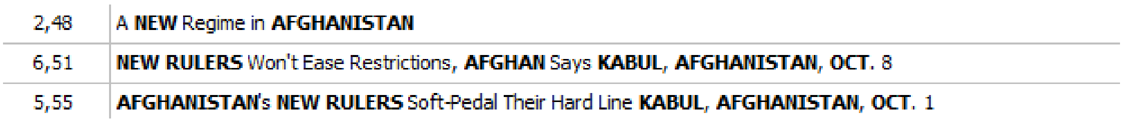
2,30	**RUSSIA** TELLS **NATO** IT ACCEPTS OFFER ON A FORMAL LINK BRUSSELS, Dec. 11
2,34	**RUSSIA**'s Concerns Aside, **NATO** Must Expand
2,44	**RUSSIA** and **NATO**

Figure 5.6. Second of three example clusters from transition between concept273 - concept16.



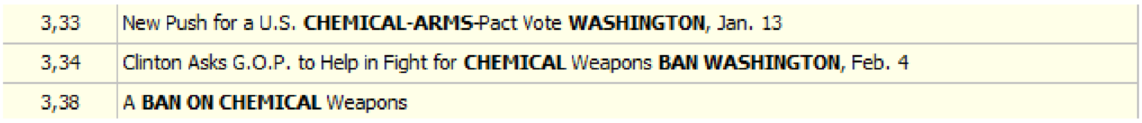
2,30	**NETANYAHU** FIGHTS REPORT BY POLICE **JERUSALEM**, April 17
2,27	'Fed Up' With Criticism, **NETANYAHU** Lashes Out **JERUSALEM**, March 12
2,26	**NETANYAHU** Can't Avoid Oslo **PEACE** Framework
3,30	King Hussein Rebukes **NETANYAHU** For 'Intent to Destroy' **PEACE** Plan **JERUSALEM**, March 11
2,40	**NETANYAHU**'s **PEACE** Plan

Figure 5.7. Third of three example clusters from transition between concept273 - concept16.



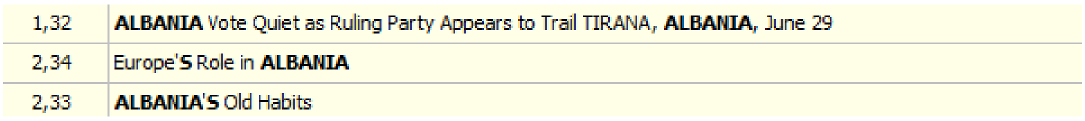
2,48	A **NEW** Regime in **AFGHANISTAN**
6,51	**NEW RULERS** Won't Ease Restrictions, **AFGHAN** Says **KABUL**, **AFGHANISTAN**, **OCT**. 8
5,55	**AFGHANISTAN**'s **NEW RULERS** Soft-Pedal Their Hard Line **KABUL**, **AFGHANISTAN**, **OCT**. 1

Figure 5.8. First of two example clusters from transition between concept397 - concept273.



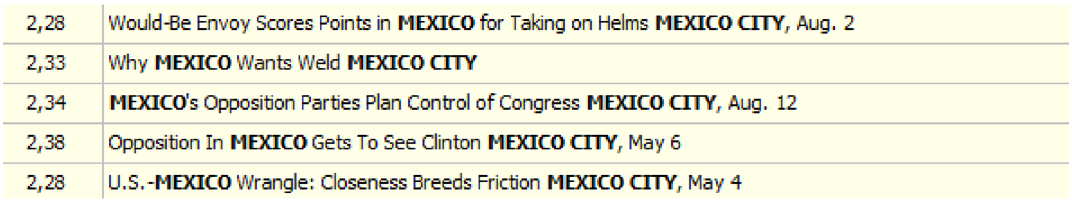
3,33	New Push for a U.S. **CHEMICAL-ARMS**-Pact Vote **WASHINGTON**, Jan. 13
3,34	Clinton Asks G.O.P. to Help in Fight for **CHEMICAL** Weapons **BAN WASHINGTON**, Feb. 4
3,38	A **BAN ON CHEMICAL** Weapons

Figure 5.9. Second of two example clusters from transition between concept397 - concept273.



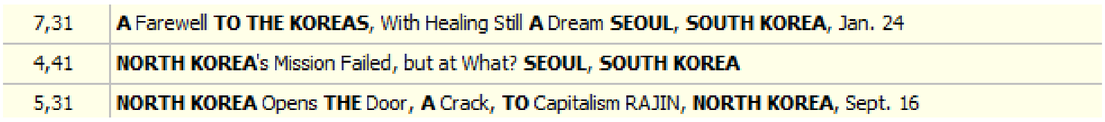
1,32	**ALBANIA** Vote Quiet as Ruling Party Appears to Trail TIRANA, **ALBANIA**, June 29
2,34	Europe'**S** Role in **ALBANIA**
2,33	**ALBANIA'S** Old Habits

Figure 5.10. First of two example clusters from transition between concept428 - concept397.



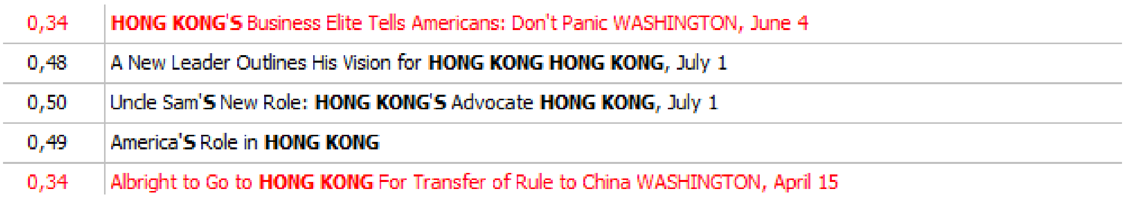
2,28	Would-Be Envoy Scores Points in **MEXICO** for Taking on Helms **MEXICO CITY**, Aug. 2
2,33	Why **MEXICO** Wants Weld **MEXICO CITY**
2,34	**MEXICO**'s Opposition Parties Plan Control of Congress **MEXICO CITY**, Aug. 12
2,38	Opposition In **MEXICO** Gets To See Clinton **MEXICO CITY**, May 6
2,28	U.S.-**MEXICO** Wrangle: Closeness Breeds Friction **MEXICO CITY**, May 4

Figure 5.11. Second of two example clusters from transition between concept428 - concept397.



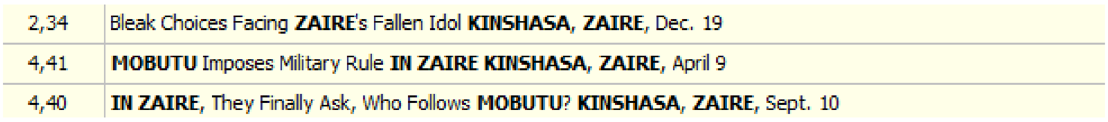
7,31	**A** Farewell **TO THE KOREAS**, With Healing Still **A** Dream **SEOUL**, **SOUTH KOREA**, Jan. 24
4,41	**NORTH KOREA**'s Mission Failed, but at What? **SEOUL**, **SOUTH KOREA**
5,31	**NORTH KOREA** Opens **THE** Door, **A** Crack, **TO** Capitalism RAJIN, **NORTH KOREA**, Sept. 16

Figure 5.12. First of two example clusters from transition between concept604 - concept428.



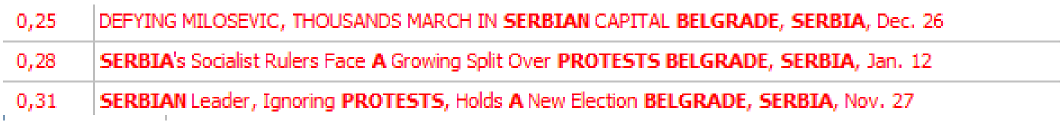
0,34	**HONG KONG'S** Business Elite Tells Americans: Don't Panic WASHINGTON, June 4
0,48	A New Leader Outlines His Vision for **HONG KONG HONG KONG**, July 1
0,50	Uncle Sam'**S** New Role: **HONG KONG'S** Advocate **HONG KONG**, July 1
0,49	America'**S** Role in **HONG KONG**
0,34	Albright to Go to **HONG KONG** For Transfer of Rule to China WASHINGTON, April 15

Figure 5.13. Second of two example clusters from transition between concept604 - concept428.



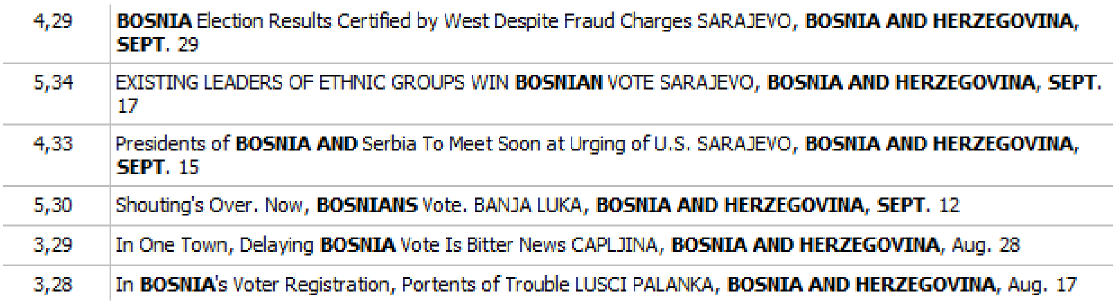
2,34	Bleak Choices Facing **ZAIRE**'s Fallen Idol **KINSHASA**, **ZAIRE**, Dec. 19
4,41	**MOBUTU** Imposes Military Rule **IN ZAIRE KINSHASA**, **ZAIRE**, April 9
4,40	**IN ZAIRE**, They Finally Ask, Who Follows **MOBUTU**? **KINSHASA**, **ZAIRE**, Sept. 10

Figure 5.14. Example cluster from transition between concept631 - concept604.



0,25	DEFYING MILOSEVIC, THOUSANDS MARCH IN **SERBIAN** CAPITAL **BELGRADE**, **SERBIA**, Dec. 26
0,28	**SERBIA**'s Socialist Rulers Face **A** Growing Split Over **PROTESTS BELGRADE**, **SERBIA**, Jan. 12
0,31	**SERBIAN** Leader, Ignoring **PROTESTS**, Holds **A** New Election **BELGRADE**, **SERBIA**, Nov. 27

Figure 5.15. Example cluster from transition between concept632 - concept631.



4,29	**BOSNIA** Election Results Certified by West Despite Fraud Charges SARAJEVO, **BOSNIA AND HERZEGOVINA**, **SEPT**. 29
5,34	EXISTING LEADERS OF ETHNIC GROUPS WIN **BOSNIAN** VOTE SARAJEVO, **BOSNIA AND HERZEGOVINA**, **SEPT**. 17
4,33	Presidents of **BOSNIA AND** Serbia To Meet Soon at Urging of U.S. SARAJEVO, **BOSNIA AND HERZEGOVINA**, **SEPT**. 15
5,30	Shouting's Over. Now, **BOSNIANS** Vote. BANJA LUKA, **BOSNIA AND HERZEGOVINA**, **SEPT**. 12
3,29	In One Town, Delaying **BOSNIA** Vote Is Bitter News CAPLJINA, **BOSNIA AND HERZEGOVINA**, Aug. 28
3,28	In **BOSNIA**'s Voter Registration, Portents of Trouble LUSCI PALANKA, **BOSNIA AND HERZEGOVINA**, Aug. 17

Figure 5.16. Example cluster from transition between concept1024 - concept641.

CHAPTER 6

CONCLUSION

Amount of news articles archived by news companies urges the creation of an automated system for news chain production to support investigative journalists and intelligence analysts. Within the scope of this study, an automatization is proposed for news chain creation based on lattice structure.

To be precise, the main objective of this study is:

To check the utility of partial order relations specifically concept lattices of news articles out of the inverted index structures in automating the construction of news chains.

In order to fulfill this requirement; selected keywords, which are related with news chain themes, are transformed into a lattice structure. It appears that the constructed lattice structure needs to be reduced or particular paths should be selected in generating candidate news chains.

We performed a case study to validate the potential of a particular path of the concept lattice in generating candidate news chains. To be specific, we go through a path by reducing the extent set one by one. Then, examining the intent differences between concepts of the lattice along the path, candidate news chains are produced.

Although promising news chains can be observed in the set, a systematic manual check is required to test whether those are good (coherent) ones or not. There is also a linear programming solution for news chain coherence check proposed by Shahaf and Guestrin (2010), implemented by Ozkahraman (2015), which will be used to check coherence of news chains and be compared with results of manual check. Using this tool, we will evaluate the success of the proposed method more systematically.

One more extension to the current work is applying the methodology to this dataset with different seed keyword sets and test the results. News articles are produced based on closed concepts. Therefore, changing keyword combination should produce different closed concepts and different news chains.

REFERENCES

Ahmed, A., Q. Ho, J. Eisenstein, E. Xing, A. J. Smola, and C. H. Teo (2011). Unified analysis of streaming news. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, New York, NY, USA, pp. 267–276. ACM.

Allan, J., J. Carbonell, G. Doddington, J. Yamron, Y. Yang, J. A. Umass, B. A. Cmu, D. B. Cmu, A. B. Cmu, R. B. Cmu, I. C. Dragon, G. D. Darpa, A. H. Cmu, J. L. Cmu, V. L. Umass, X. L. Cmu, S. L. Dragon, P. V. M. Dragon, R. P. Umass, T. P. Cmu, J. P. Umass, and M. S. Umass (1998). Topic detection and tracking pilot study final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218.

Border, K. C. (2011). Preliminary notes on lattices. `http://people.hss.caltech.edu/~kcb/Notes/Lattice.pdf`. Accessed: 2014-09-03.

Brüggemann, R. and G. Patil (2011). *Ranking and Prioritization for Multi-indicator Systems: Introduction to Partial Order Applications*. Environmental and Ecological Statistics. Springer New York.

Gallier, J. (2014). *Discrete Mathematics (Second Ed.)*. Springer-Verlag New York.

Ganter, B., G. Stumme, and R. Wille (2005). *Formal concept analysis: foundations and applications*, Volume 3626. Springer Science & Business Media.

Gillenwater, J., A. Kulesza, and B. Taskar (2012). Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, Stroudsburg, PA, USA, pp. 710–720. Association for Computational Linguistics.

Hossain, M. S., C. Andrews, N. Ramakrishnan, and C. North (2011). Helping intelligence analysts make connections. In *Scalable Integration of Analytics and Visualization*.

Hossain, M. S., P. Butler, A. P. Boedihardjo, and N. Ramakrishnan (2012). Storytelling in entity networks to support intelligence analysts. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, New York, NY, USA, pp. 1375–1383. ACM.

Knuth, D. E. (1998). *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc.

Kowalski, G. (1997). *Information Retrieval Systems: Theory and Implementation* (1st ed.). Norwell, MA, USA: Kluwer Academic Publishers.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Myat, N. and K. Hla (2005, Sept). A combined approach of formal concept analysis and text mining for concept based document clustering. In *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*, pp. 330–333.

Ozkahraman, O. (2015, June). Linking dots together. Bachelor thesis, Izmir Institute of Technology.

Porter, M. F. (1997). Readings in information retrieval. Chapter An Algorithm for Suffix Stripping, pp. 313–316. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rose, T., M. Stevenson, and M. Whitehead (2002). The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *In Proceedings of the Third International Conference on Language Resources and Evaluation*, pp. 29–31.

Sandhauts, E. (2008). The new york times annotated corpus overview. `https://catalog.ldc.upenn.edu/docs/LDC2008T19/new_york_times_annotated_corpus.pdf`. Accessed: 2014-09-16.

Shahaf, D. and C. Guestrin (2010). Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discov-*

ery and Data Mining, KDD '10, New York, NY, USA, pp. 623–632. ACM.

Shahaf, D., C. Guestrin, and E. Horvitz (2012a). Metro maps of science. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, New York, NY, USA, pp. 1122–1130. ACM.

Shahaf, D., C. Guestrin, and E. Horvitz (2012b). Trains of thought: Generating information maps. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, New York, NY, USA, pp. 899–908. ACM.

Valtchev, P., D. Grosser, and C. Roume (2003). Galicia: an open platform for lattices. In *Using Conceptual Structures: Contributions to the 11th Intl. Conference on Conceptual Structures (ICCS'03)*, pp. 241–254.

Wang, D., T. Li, and M. Ogihara (2012). Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs.

Wille, R. (2009). Restructuring lattice theory: An approach based on hierarchies of concepts. In *Proceedings of the 7th International Conference on Formal Concept Analysis*, ICFCA '09, Berlin, Heidelberg, pp. 314–339. Springer-Verlag.

Yevtushenko, S. (2004, October). *Computing and Visualizing Concept Lattices*. Ph. D. thesis, TU Darmstadt.

Zaki, M. J. and C.-J. Hsiao (2002). Charm: An efficient algorithm for closed itemset mining. *SDM 2*, 457–473.

Zaki, M. J. and C.-J. Hsiao (2005, April). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Trans. on Knowl. and Data Eng. 17*(4), 462–478.

Zaki, M. J. and N. Ramakrishnan (2005). Reasoning about sets using redescription mining. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, New York, NY, USA, pp. 364–373. ACM.

APPENDIX A

MATERIAL AND METHOD

Selected Topics of Reuters Corpus Volume 1

- CURRENT NEWS - POLITICS
- STRATEGY/PLANS
- INTERNAL POLITICS
- INTERNATIONAL RELATIONS
- DOMESTIC POLITICS
- INTERNATIONAL RELATIONS
- EUROPEAN COMMUNITY
- EC EXTERNAL RELATIONS
- EC GENERAL
- DEFENCE
- WAR, CIVIL WAR

Selected Topics of New York Times Annotated Corpus

- International Relations
- Iraq War (2003-)
- Politics and Government
- Terrorism
- United States International Relations
- United States Politics and Government

- Condominiums
- Conservatism (US Politics)
- Defense and Military Forces
- Embargoes and Economic Sanctions
- Espionage
- Extradition
- Foreign Aid
- Foreign Investments
- Freedom and Human Rights
- Governors (US)
- Gun Control
- Kurds
- Lobbying and Lobbyists
- Military Bases and Installations
- Missiles and Missile Defense Systems
- Nuclear Weapons
- Palestinians
- Political Advertising
- Political Prisoners
- September 11 (2001)
- Sunni Muslims
- United States Defense and Military Forces
- War Crimes, Genocide and Crimes Against Humanity

APPENDIX B

RESULTS

Table B.1.: Most frequent 100 keywords of politics sub-dataset of New York Times Annotated Corpus

Order	Keyword	Frequency	Order	Keyword	Frequency
1	govern	11266	25	call	4375
2	presid	10278	26	china	4199
3	offici	9069	27	republican	4133
4	american	8838	28	militari	4048
5	polit	8707	29	month	4044
6	unit	8623	30	public	3988
7	clinton	7888	31	minist	3974
8	peopl	7430	32	law	3962
9	nation	7423	33	support	3957
10	parti	7101	34	issu	3926
11	time	6491	35	power	3910
12	countri	6397	36	foreign	3812
13	elect	5868	37	war	3781
14	leader	5567	38	isra	3741
15	democrat	5136	39	includ	3695
16	citi	4880	40	secur	3668
17	hous	4810	41	administr	3627
18	report	4626	42	meet	3439
19	day	4615	43	plan	3406
20	forc	4606	44	percent	3223
21	week	4566	45	senat	3167
22	palestinian	4432	46	world	3152
23	campaign	4395	47	washington	3116
24	offic	4380	48	polici	3069

Order	Keyword	Frequency	Order	Keyword	Frequency
49	vote	3051	75	congress	2394
50	million	3029	76	question	2391
51	peac	2972	77	court	2386
52	recent	2959	78	aid	2381
53	polic	2874	79	kill	2365
54	white	2833	80	york	2341
55	bomb	2810	81	effort	2332
56	feder	2807	82	author	2324
57	control	2806	83	netanyahu	2305
58	compani	2800	84	run	2278
59	intern	2766	85	critic	2261
60	talk	2721	86	move	2191
61	israel	2718	87	term	2155
62	ms	2699	88	ago	2149
63	major	2668	89	remain	2145
64	rule	2655	90	famili	2144
65	organ	2619	91	charg	2141
66	econom	2596	92	opposit	2141
67	right	2520	93	continu	2140
68	investig	2518	94	decis	2114
69	build	2487	95	told	2094
70	live	2481	96	home	2079
71	attack	2467	97	candid	2076
72	busi	2426	98	news	2067
73	prime	2415	99	trade	2060
74	money	2402	100	depart	2032