T.C. DOGUS UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

COMPUTER AND INFORMATION SCIENCES MASTER PROGRAM

AUTOMATIC HYPERLINK GENERATION

M.S. Thesis

Deniz ŞERİFOĞLU
200791002

Advisor: Prof. Dr. Selim AKYOKUŞ

Istanbul, September 2010

T.C. DOGUS UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

COMPUTER AND INFORMATION SCIENCES MASTER PROGRAM

AUTOMATIC HYPERLINK GENERATION

M.S. Thesis

Deniz ŞERİFOĞLU
200791002

Advisor: Prof. Dr. Selim AKYOKUŞ

Istanbul, September 2010

ACKNOWLEDGEMENT

The objective of this thesis is to develop a program that generates the links automatically. I have chosen this topic because nowadays there aren't enough examples about this subject in Turkish.

This study was spread to a long period of time and it has finally finished.

Special thanks to Prof. Dr. Selim Akyokuş for his guidance, his effort in advising to me and his appreciation of my work and his belief on me.

I also thank to my wife Melek Şerifoğlu. She always gave me her support the whole period of project submission.

Finally I thank to my friends Kıvanç Onan and Yeliz Ekinci for their assistances.

Istanbul, September 2010 Deniz ŞERİFOĞLU

TABLE OF CONTENTS

ABSTRACT

One of the most important inventions of today is the Internet. Hundreds of millions of people anytime, from anywhere, can enter the Internet. On the Internet web pages are represented in HTML format and pages are linked through hyperlinks. Normally hyperlinks are defined by users manually. The objective of this thesis is to design a system that generates hyperlinks automatically.

For this objective, a robot called SeaGEN has been developed. The robot SeaGEN analyses a web page and generates hyperlinks for certain words/phrases.

During this Master study data mining techniques were used to generate hyperlinks. A classification system was developed. A training dataset was collected from Vikipedi. This training set was used for training the classification system. WEKA open source data mining software program was used for classification. The trained classification system generates hyperlinks automatically for a given set of pages.

ÖZET

Günümüzün en önemli buluşlarından biri hiç şüphesiz internettir. Her an dünyanın herhangi bir yerinden yüzlerce milyon kişi internete girebilmektedir. İnternet ortamında ağ sayfaları HTML formatında gösterilmekte ve ağ sayfalarına erişim hiperlinklerle gerçekleşmektedir. Genel olarak hiperlinkler kullanıcılar tarafından el ile manüel olarak tanımlanırlar. Bu tezin amacı hiperlinkleri otomatik olarak oluşturacak bir sistem tasarlamaktır.

Bu amacı gerçekleştirmek için SeaGEN isimli bir robot geliştirilmiştir. SeaGEN ağ sayfalarını analiz ederek çeşitli kelime ve deyimler için hiperlinkler oluşturmaktadır.

Bu çalışmada hiperlinkler oluşturmak için veri madenciliği tekniklerinden yararlanılmıştır. Önce bir sınıflandırma sistemi geliştirilmiştir. Vikipediden deneme amaçlı bir veri seti alınmıştır. Sınıflandırma sistemi deneme seti üzerinde test edilmiştir. Bu sınıflandırma için WEKA isimli, açık kaynak kodlu bir veri madenciliği yazılımı kullanılmıştır. Uygulama sonunda, test edilen sınıflandırma sisteminin, belirlenen ağ sayfaları için hiperlinkleri otomatik olarak oluşturduğu tespit edilmiştir.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

NLP	Natural Language Processing
API	Application Programming Interface
GUI	Graphical User Interface
HTML	Hypertext Markup Language
HTTP	Hypertext Transport Protocol
IT	Information Technology
IR	Information Retrieval
VSM	Vector Space Model
SQL	Structured Query Language
JDBC	Java Database Connectivity
URL	Uniform Resource Locator
URI	Uniform Resource Identifier
DCM	Document Cluster Map
WCM	Word Cluster Map
WWW	World Wide Web
VM	Virtual Machine
AI	Artificial Intellgince
SE	Search Engine
SOM	Self Organizing Map
k-NN	k-Nearest Neighbor

1. INTRODUCTION

The usage of hypertexts formats has been widely distinguished and admitted. The access mechanism provided by HyperText Markup Language (HTML) is one of the reasons of the growth of the World Wide Web (WWW). It's estimated that there exist more than 3 billion web pages worldwide and more than 1 million web pages being created newly each day. HTML is the standart format for representing document in WWW. HTML provides hyperlinking mechanism that enables navigation among documents.

Although many types of document formats such as PDF and DOC are used on the internet. The absolute majority of documents are encoded using HTML. Hyperlinks among HTML encoded documents are specified manually.

To transform a normal text to a hypertext we have to determine where to put in a link in the text. A link joins between 2 written documents where at single end of the link is the root text which could be a single word or a group of words and at the other ending is the goal text which could be a different written document or another location of the very same document. Different cases of links could be applied depending to the forms of functionality that require to be applied by the hypertext. According to Agosti, Crestani, and Melucci (1997), there are 3 types of links, these are structural links, referential links, and associative links. Structural links are the links which the information contained within the hypermedia web applications are typically organised in some suitable fashion. Referential links provide the links between an item of information and an elaboration or explanation of that information. And associative links are the instantiation of a semantic relationship between information elements. The first 2 types of links are generally denotative and might be easily produced manually or automatically.

Automatic hyperlink generation is one of the current research topics. That is not studied much in nowadays. There is little research literature on this area. Automatic hyperlink generation consists of two steps. First one is to decide on root text to be linked and second is to decide on documents to be linked to the root document.

The objective of this thesis is to develop a system to generate hyperlinks automatically. On this study first a web crawler is developed to collect training web pages from WWW.

Trained web pages are collected from Vikipedi. These pages are lexically analyzed and parsed. After the analysis, stemming and stopwords removal preprocessing techniques are applied. Then we obtained a document term matrix. Then this document term matrix is used as an input in a classifier. As a classifier we used WEKA open source data mining software program. k-NN algorithm used as a classifier. This classifier is then used to analyze several test page in order generate automatic hyperlinks.

This thesis first overviews web mining and text mining in chapter 2 & 3. In chapter 4, preprocessing methods for text processing and Zemberek stemming library are introduced. In chapter 5, previous related research work are on about automatic hyperlink generation is summarized. Chapter 6 describes the developed automatic hyperlink generation system, called SeaGEN. Chapter 7 includes the conclusion of the thesis.

2. WEB MINING

Web mining is the application program of data mining techniques to distinguish models from the internet. According to analysis marks, web mining could be separated into 3 divergent cases, which are Web usage mining,Web content mining and Web structure mining.

2.1 Web usage mining

Web usage mining is the operation from determining what users are searching on the Internet. A few users could be viewing just textual information, whereas just about other people could be occupied in multimedia system information.

2.2 Web structure mining

Web structure mining is the method of utilising graph theory to study the node and association construction of an internet site. According to the case of web structural data, web structure mining could be separated into two forms:

1. Educing forms from links in the internet: a link is a functional element that joins the web page to another placement.

2. Mining the written document structure: analysis of the tree-shaped structure of page structures to identify HTML or XML chase utilisation.

2.3 Web Content Mining

Web Content Mining identifies the automatic research of data resource accessible online [Madria 1999], and implies mining web data content. Opposed to Web Usage Mining or Web Structure Mining, Web Content Mining accent about the articles of the web page just the hyperlinks.

The Web content mining are severalized from two unlike viewpoints: Information Retrieval View and Database View. R. Kosala summed up the enquiry acts gone ambiguous information and semi-structured information from information retrieval view. It demonstrates that most of the enquiries employment bag of words, which is supported

the statistics approximately exclusive words in closing off, to comprise ambiguous text and admit separate word determined in the training principal because characteristics. As the semi-structured information, whole the acts apply the HTML structures interiors the paperses and a few applied the hyperlink construction between the documents for document representation. Because as the database aspect, called for to accept the best data direction and questioning about the Web, the mining all of the time efforts to deduce the structure of the internet site of to translate a internet site to convert a database.

An in-depth review of the enquiry about the application program of the methods from machine acquiring, statistical figure acknowledgment, and data mining to examining hypertext is allowed by S. Chakrabarti. It's a beneficial imagination to comprise mindful of the late encourages in content mining enquiry.

2.4 Web mining Pros and Cons

Pros

Web mining fundamentally has got several advantages which forms this technology engaging to corps admitting the government agencies. This technology has enabled ecommerce to make individualised merchandising, which finally effects in greater deal bulks. The government agencies are utilising this technology to classify threats and battle against act of terrorism. The anticipating potentiality of the mining application program could profits the society with describing outlaw actions. The societies could demonstrate finer customer kinship along establishing them incisively what they demand. Societies could empathise the demands from the customer more well and they could respond to customer demands quicker. The companies could observe, appeal and keep customers; they could keep upon output prices along applying the developed in view of customer necessities. They could gain profitableness along aim pricing supported the visibilities produced. They could regular detect the customer who could nonremittal to a contender the company would attempt to hold the customer of allowing promotional passes to the particular customer, hence coming down the hazard from missing a customer or customers.

Cons

Web mining, itself, does not produce consequences, but this technology once applied upon information of individual nature could drive vexations. The most important point affecting web mining is the secrecy. Concealment is reasoned missed while data occupying an individual is incurred, applied, or distributed, particularly whenever this happens without their cognition or accept. The incurred information will be canvassed, and bunched to cast visibilities; the information will be created unknown ahead bunching and so that there are none individual visibilities. Hence these applications de-individualize the users of adjudicating them by their mouse clicks. Generalization, could be settled as an inclination by adjudicating and dealing inhabit with the base by grouping features besides on their own individual features and deserves.

Another significant refer is that the companies pulling in the information for a particular aim could expend the information for a wholly dissimilar aim, and this fundamentally infracts the user's concerns. The developing style of trading individual information as a goods advances internet site proprietors to deal individual information got from their internet site. This style has expanded the number of information being caught and listed growing the likelihood of one's privateness being occupied. The companies which bargain the information are obligated make it unknown and these companies are believed sources of some particular dismission of mining patterns. They're lawfully obligated for the capacities of the expiration; some inaccuracies in the expiration will effect in grievous causes, but there's no more jurisprudence keeping them from switching the data.

Some mining algorithms could apply contentious properties as if gender, race, faith, or intersexual preference to categorise humans. These exercises could be against the anti-discrimination lawmaking. The applications pull through difficult to describe the apply of such disputable properties, and there's no heavy formula against the utilisation of such algorithms with such properties. This action might effect in abnegation of service or a favour to an personal supported his race, faith or intimate preference, right now this position could be annulled along the high honorable criteria exerted along the data mining company. The gathered information is being created anonymous so that, the received data and the received designs can't be delineated back to an individual. It could appear as

though this airses no menace to one’s concealment, in reality a lot additional data could be derived by the application by aggregating 2 apart dishonest information from the user.

Web mining is the method of examining an categorisation of behavioural, demographic, life-style, transactional, Internet and geographical entropy for the personalization of extends to online consumers in real time. It's the usage of artificial intelligence algorithms conjugated on information meshes for the analysis of all of this data – via software factors which mine these data sets at their first position of computer storage triggering pointed extends for consumer effects happen. In web mining there's no latent period between analysis and activity – alternatively it's one separate merged continuous process – alleviated along the apply of strategically laid software agents to mine and example multiple data sets consorted to online consumers.

Web mining could affect whole of the conventional data mining actions of categorisation, segmentation, clustering, association, prediction, and modeling the lone departure is that the analyses issue in contiguous activity. Contrary to data mining, web mining is contingent the use of software agent to trigger aimed extends for cases occur immediately. An agent is a computer program that gets activity upon behalf of a method, which in that cause could be hybrid and up selling, customer retention, risk assessment, fraud detection and counterterrorism.

Web mining confirms the power to market to individual customers supported the information gathered horizontally over multiple data sources about them separately and leveraging it instantly for cases or proceedings happen, such as once an electronic mail is obtained or a visit is to a internet site occurs. In web mining real-time data analytics gets a proceeding and reiterative method in which business determinations and activity are ceaselessly refined and formed over time in order to extend crucial ads, production, services and content to online consumers.

The information parts for web mining could imply user-provided information, server log files, cookies, form-generated datasets, email, as well as commercial demographics, life style information, previous browsing action, anterior sales, transactional data, Internet geolocation, search keywords, re-directs or referrals, and other consumer associated behaviour. In web mining reactions and extends are aimed along real time cases and

interactions. Web mining different data mining which subsisted preceding the Internet burst affects afresh prototype of data collection, integration and analysis. Web mining needs design identification via an unlined pullulates of action occurring across a determination network and not a stable storage warehouse. Web mining forms by doing data analysis via networks, using software agents to mine, collaborate and discover specifies and characteristics which could guide to increments in sales, cross-selling chances and the aiming of particular products or services.

Web mining allows an initiative the incorporated instruments for analysing whole case of data sources from multiple sections, from different locations from different arranges for an classification of deliverables specified tendency to leverage marks, risk marks for dupery, prediction of customer behavior or the innovation of customer classes or groups. Web mining enables an endeavor to purchase their routine communicatings on actionable cognition breakthrough, real-time business intelligence and placed customer responses. Web mining enables an endeavour to cause the right offer, to the right customer, as cases take place instantly.

2.5 Link Structures

Different types of hyperlinks have been specified contingent the functionalities that involve to be applied by the hypertext. Most hypertexts are assembled building use of the following 3 types of links:

2.5.1 Structural links: connect nodes of the hypertext that are colligated by the structure of the document itself. Examples of this type of link are hyperlinks associating a chapter on the chapter that comes after it, or a table of contents with each part described in it. Whenever the transformation from a edict document that's a record to hypertext is built, all the chapter/segment/subdivision functional associations could be depicted in the hypertext using structural hyperlinks. Each structure (for example. tree, graph) could be generated using structural hyperlinks.

2.5.2 Referential links: are settled with some kind of mention the writer of the original document has used. A typical hyperlink of this sort is the link that applies a reference between the origin document and a document that is cited by it.

2.5.3 Associative links: represent indefinable associatory connections between nodes. They're assembled attaining use of content-based connections between shards of text of the same document or documents of the same collection. All those hyperlinks are formed expressly accessible to the user through the hypertext network. Another type of link, the aggregated hyperlink, is built available only in a few hypertext systems, where the accumulation generalisation mechanism is configured and applied to give the hypertext couturier the theory of combining nodes that together form a new kind of node; this mechanism has been made useable in semantic information examples (Schiel, 1989) but it's stil rarely acquirable in functional database direction and hypertext systems (Smith & Smith, 1977).

Another categorisation of hyperlinks is very useful to add, that of explicit/implicit link:

• an *explicit link* is a hyperlink that attains usable an explicit reference between 2 nodes; explicit hyperlinks are built on the authoring action and they comprise the primary division of the hypertext network;

• an *implicit link* is a hyperlink that's implicitly gift in a node. An implicit hyperlink could be excited using a word deliver in a node. For example, whenever a user demands to look all nodes that arrest an exceptional word, all nodes comprising that word could be constituted useable to the user, and these implicit hyperlinks are made at run time. This means that an implicit hyperlink doesn't link a pair of nodes, but it's implicitly present in the node text and it's made and built available at run time. An example of this process is delineated in (Aalbersberg, 1992).

The network of hyperlinks comprises the only structure which could be used to navigate the hypertext. In order to navigate the hypertext, the user involves a tool capable to follow the hyperlinks: this capability is generally supplied by a tool named a web browser. A web browser commonly comprises both navigation and browsing facilities. Whenever a hyperlink doesn't subsist between 2 nodes which are semantically connected, they can't be watched (regained) by a user who's browsing the hypertext. The only way in which 2 or more related nodes that haven't been expressly colligated by hyperlinks could be thought it's by looking for the network for some string of words, keyword or property value which nodes have to portion, and this take advantage of an implicit hyperlink

between nodes. Usually it's only one particular and accurate string, keyword or property value that could be applied for inquisitory in such conditions. Most present hypertext browsing tools of complete hypertext systems can't generally provide exact match retrieval techniques that use a query language supported Boolean algebra, and that are available in most of the present operational information retrieval systems. Contrarily, the web is a big hypertext of fairly convoluted SEs, though they're only mistily comprised into the web browser themselves.

3. TEXT MINING

Text mining has been a popular subject in recent years. Numerous researchers and practicians have used different techniques in relevant studies about this subject.(Tan, 1999). Amongst these advances, the self-organizing map (SOM) (Kohonen, 1997) pattern acts an significant character. Bunches of studies had exploited SOM to cluster big aggregation of text documents. Examples could be got hold in Kaski, Honkela, Lagus, and Kohonen (1998), Lee and Yang (1999), and Rauber and Merkl (1999). Nevertheless, there's a few study had practised text mining approaches, especially the SOM advance, to automatic hypertext expression. Unitary ending study by Rizzo, Allegra, and Fulantelli (1998) used the SOM to cluster hypertext documents. Nevertheless, their work was applied for synergistic browsing and written document researching, instead of hypertext authoring.

Text mining demands the application program of methods from fields such as information retrieval, NLP, data extraction and data mining. These versatile levels of a text-mining action could be conjunctive together into a individual work flow.

IR systems describe the documents according to the IR mapping to a user's query. The most known IR systems are SEs such as Google, which describe those documents upon the WWW that are related to a set of given words. IR systems are frequently applied in libraries, where the documents are commonly not the articles themselves but digital records carrying data around the articles. These are still transferring on the advent of digital libraries, where the documents being retrieved are digital editions of articles and journals.

IR systems provide us to specialise the set of documents that are crucial to a specific problem. Because text mining requires holding identical computationally-intensive algorithms to big written document ingatherings, IR could accelerate the analysis substantially by coming down the amount of documents for analysis. For instance, whenever we're occupied in mining data only if approximately protein interactions, we could bound our analysis to documents that carry the key of a protein, or some phase of the verb 'to interact' or one of its synonyms.

4. PREPROCESSING STEPS

4.1 Lexical Analysis

Lexical analysis is the method of changing a succession of cases into a sequence of tokens in computer science. A computer program or procedure which executes lexical analysis is named a lexical analyzer, lexer or scanner. A lexical analyzer frequently subsists as an individual procedure which is known as a parser or a different procedure.

A token is a string of parts, classified harmonising to the conventions for a symbolisation (e.g. IDENTIFIER, TAG, COMMA, NUMBER, etc.). The operation from organising tokens from an input file of terms is known as tokenization and the lexer categorises them concording to a symbol case. A token could look like anything that is valuable for processing an input text file.

4.2 Stemming

In most cases, morphological forms of words have similar semantic renditions and could be believed as equivalent for the aim of IR applications. For this reason, a list of alleged stemming Algorithms, or stemmers, have been trained, which effort to abridge a word to its stem or root form. Hence, the key terms of a query or document are presented by stems instead of by the original words. This not only implies that different variances of a term could be coalesced to a individual illustration phase it also comes down the lexicon size, that's, the number of distinct terms wanted for constituting a set of documents. A smaller lexicon size issues in a keeping of computer storage space and processing time.

For IR aims, it does not typically matter whether the stems generated are genuine words or not – hence, "computing" could be stemmed to "comput" allowed that (another words with the same 'base meaning' are coalesced to the same shape, and (b) words with different meanings are held separate. An algorithm which efforts to commute a word to its lingually castigate root ("compute" in this case) is occasionally named a lemmatiser.

Examples of products using stemming algorithms would be SEs such as Lycos and Google, and in addition to thesauruses and other products using NLP for the purpose of IR. Stemmers and lemmatizers also have applications more widely inside the area of Computational Linguistics.

4.2.1 Zemberek

Zemberek is an open source NLP library for Turkish Language. This is the first open source library in this area. Project was once called "Tspell" in java.net. Program will try to puzzle out subjects about language such as spell checking, word and sentence analysis and many another lots of things.

4.2.1.1 How Zemberek works?

Dictionary & Stem-tree

How Zemberek determine a word is Turkish or not? The answer is “If we could separate the word into its stem and additionals, it might Turkish or it might not.” Briefly to empathise the word is Turkish or not we could do morphological analysis. Turkish spelling in the ancient to be able to assign the most often used words in a data file from the data file to agree if the words when it seem logical at first sight it looks that path but a little unfunctional as the review was also believed. 98-99% truths on this sort of techniques are insufficient still for the exercise must hold a million words. Zemberek with a identical easy system of rules to analyse the morphologic word. Candidates might be given prior to the origin of the word sets, and then that might be appropriate for attachment to the root in order to add this study. During this action we also take admittance the word, and so supplied the suitable origin of the word is Turkish, and as well found the intends, if we origin for any of the expected consequence isn't yet available, and then the word is Turkish. The initial step in this technique when tests the method of getting the origins candidate. 1st candidate for getting the origins whole the roots of the words must be detected in Turkish. Turkey Turkish roots on spring in a 30,000 pack on it a point, this guide to whole characters of origins, and in particular examples are marked concording to. The former for the Turkish language executions ought to expect a synonymous origin lexicon.

Getting the root of the word applied to candidates Zemberek this stem would order exceptional tree, recognition of candidates on this specific tree could be created very rapidly. In that tree in the roots of word of are identified accordant to the articles. For

instance, in the example at a lower place, "BAZ" of the stem, severally B-A-Z-labeled nodes linked to the last point en route.

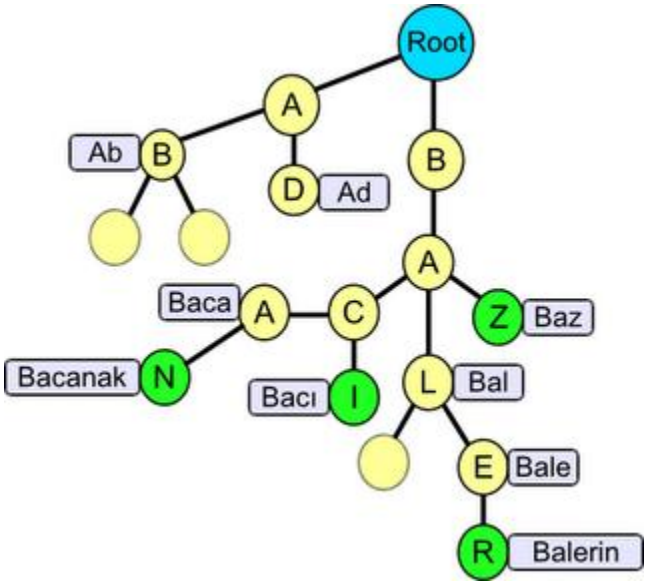

Figure 4.1 Analysis of Word “Balerin” ([19])

Tree stems from root nodes and hyperlinks that belong to nodes, once the “Kitap” derived for another buffering is likely to have converted whenever the origins are also expanded to the tree roots. Whenever this has modified, but also argues the master stem.

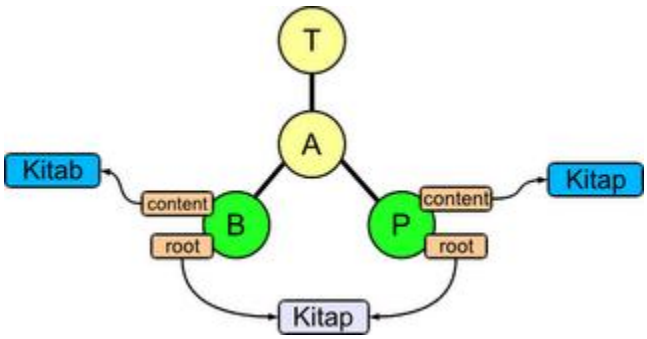

Figure 4.2 Analysis of Word “Kitap” ([19])

Some identical good origins of the tree foundation method wants acquiring into account extremum events, producing a more beneficial computer memory efficiency and performance from the tree identified above, concording to the formulas to conduct

differently for the next forethoughts ought to have been 7-8 events.

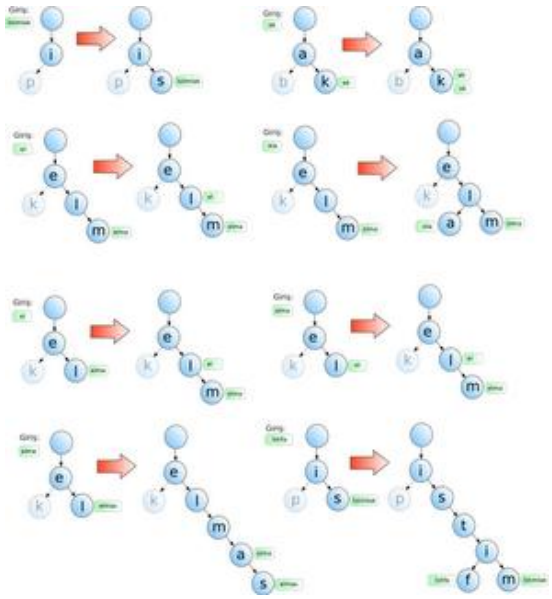

Figure 4.3 Analysis of Words (http://code.google.com/p/zemberek/)

4.3 Stop Words Removal

Stopwords are often applied, contex free grammar in a language such as prepositions, numerals, pronouns, conjunctions, stereotyped abbreviations etc.

Stopwords are general words that carry more insignificant meaning than keywords. Generally SEs take away stopwords from a keyword expression to generate the most crucial solution. Stopwords aim much less traffic than keywords.

The articles could hold up to 70% of stopwords, only 30% of words are keywords that efforts SE traffic.

The stopwords list in Turkish language is shown below:

5. RELATED WORK

Hsin-Chang Yang, Chung-Hong Lee worked in the article “A text mining approach for automatic construction of hypertexts” that the enquiry about automated hypertext building comes out quickly in the final ten as at that place subsists an imperative demand to transform the large quantity of bequest written document into internet pages.

In that study, they'd advise afresh automated hypertext building technique supported a text mining advance. Their technique holds the self-organizing map algorithm to bunch several at text documents inward a preparing principal and yield 2 maps. And then they apply these maps to discover the origins and goals from about significant links inside these developing documents. The built links are then entered into the educating written document to interpret them into hypertext shape. Such as transformed text file could organise the novel corpus. Entering paperses could besides be interpreted into hypertext shape and expanded to the principal by the identical advance. Their technique had been examined about an arrange from text paperses gathered from a newswire internet site. While they lone apply Chinese text paperses, their advance could be practiced to any paperses that could be metamorphosed to a band of exponent conditions.

Enquiry about automated building of hypertext grew generally from the information retrieval area. A study of the apply from IR methods for the automated hypertext building could be got hold in Agosti et al. (1997). In that respect there was no neural network settled techniques had built important donation in that area harmonising to their study. A different appraise from link genesis is submitted in Wilkinson and Smeaton (1999). Salton et al. (Salton & Buckley, 1989; Salton, Buckley, & Allan, 1992; Salton, Allan, & Buckley, 1993, 1994) apply the data rendered from the computing of the similarity between sherds from paperses systematic to describe subject hyperlinks. Their studies are founded with vector space model and supply basis as several hypertext building techniques. Agosti and Crestani (1993) excogitation a methodological analysis for automated building from hypertexts that would be applied in IR chores. They set up a conception pattern that comprises of 3 layers, that is to say index term level, document level and concept level and organize a five-steps procedure to build the hyperlinks amongst documents, index terms, and concepts. They apply general IR methods in the

above-named action to obviate justification of novel techniques. The methodological analysis is ulterior carried out in TACHIR (Agosti, Crestani, & Melucci, 1996) and Hyper-TextBook Project (Crestani & Melucci, 2003). Dalamagas and Dunlop (1997) advise a methodological analysis since the automated building of hypertext which constitutes customised to the area of newsprint archives and is supported Salton's advance. Meanders, which are substories inside a level, are described from giving clustering methods to reports' sections that represent to subtopics inside the primary subject of a report and so connecting the sections which lie to the equivalent cluster. The technique is ulterior followed through in News Hypertext System (Dalamagas, 1998). Green (1997, 2000) advises a technique as automated hypertext building from taking apart lexical chains in a text supported the Wordnet. The lexical chains are described and their grandness is appraised. Such as grandness is by and by applied in computing paragraph similarity and constructing inter-paragraph hyperlinks. Additional cases of hyperlinks, that is to say the inter-article hyperlinks, are assembled along finding out the similarity of the 2 exercise set of chains arrested in 2 reports. Kurohashi, Nagao, Sato, and Murakami (1992) build hypertext construction since a computer science lexicon from getting sentential forms amongst the texts to demonstrate the copulations between formulates. Such advance is quite constrictive and could lone be implemented to particular principals specified the one they applied. A former exercise along Salminen, Tague-Sutcliffe, and McClellan (1995) aggregates schematic grammar and papers indexing methods to change semi-structured text to hypertext. Their technique could lone practice to those paperses that have got hierarchical data structure. Shin, Nam, and Kim (1997) also apply an intercrossed advance which aggregates 2 similarity standards, that is to say the statistical similarity and semantic similarity, to produce beneficial hypertext, where the statistical similarity is conventional tf$idf burdening system and the semantic similarity underlies a synonym finder. Lin, Hamalainen, and Whinston (1996) formulate an organization mentioned Knowledge-Based HTML Document Generator that comprises the cognition descended of regularly updated databases to alleviate the foundation and sustainment of HTML documents. Nevertheless, this organization can't be held to discretional paperses that flowed arbitrarily and trusts along analytic thinking besides because authoring expertness in the origination method. Tudhope and Taylor

(1997) apply semantic familiarity amounts during content, spacial, and impermanent properties to produce links for navigation aim.

Allan (1996, 1997) talk about the direction the links may be automatically written. He computes the similarity between paperses and between divisions of paperses. The written documents, or divisions of them, are then connected allotting to their similarity. 6 characters of links, that is to say aggregate links, revision links, tangent links, summary/expansion links, comparison/contrast links and equivalence links are described along studying the hyperlinks.

They reported a scenario that the automated hypertext building action might practice. A newswire internet site made HCY News develops news reports from newsmen everywhere a nation. A newsman remands his articles in evident texts since fast printing. With no postprocessing, the HCY news interprets these reports into HTML arrange and sends them along their internet site. A lecturer browses along the news site and studies a report almost NBA games that references the player Michael Jordan. He would like to obtain a lot of reports about Michael Jordan. The simply path he could do here is to apply the research installation allowed in this internet site to look for the subject he wishes. Nonetheless, he determines there are more than a thousand reports about M. Jordan and is deluged. But then, he might also seek a topic and disappointedly gets no issue. On either manner he might determine to log out of this internet site since it's so inopportune for him to determine concerning news. Afterwards missing a lot of orbs, HCY news selects to allow navigational hyperlinks in their newspaper article. Even so, non-automatic building of these hyperlinks appears unsuitable as it demands extra individual imaginations and is easy, inappropriate, and discrepant. The precede of an automated hypertext building organization figures out whole these troubles. Abreast of coursing the arrangement more bequest newspaper article in field textual matter data format, the organization translates these reports into hypertext figure consistently briefly time. Entering news might besides be metamorphosed and integrated into the hypertext information exercise set. At once, while a user studies a report, he would see about significant conditions on links about them. He could chatter with a hyperlink while he's worried about the subject. He might in addition to carry out extra semantically concerned reports immediately along the supporter of further hyperlinks rendered from the hypertext building organization. The

operational parts and information fluxes in their technique are described in Fig. 5.1. They concisely identify the senior abuses from the building action in the following:

Step 1. *Corpus building*. An exercise set from evident texts is accumulated and would be applied since developing. This exercise set from texts would be metamorphosed to hypertexts afterwards.

Step 2. *Preprocessing*. Text preprocessing is executed to metamorphose this textual matter into indicator conditions. Versatile forms from indicant condition choice techniques are put on to come down the amount of index terms. A textual matter is then transmuted into a transmitter consorting to its index conditions.

Step 3. *Labeling and clustering*. They built a neuron map and execute the self-organizing map clustering algorithmic rule upon the papers transmitters. A tagging action is acquitted to label all papers to a few neuron. They likewise utilised a different tagging method to label a few significant conditions to the neurons. Afterwards the tagging procedures, they got 2 characteristic maps.

Step 4. *Hyperlink generation*. A patent textual matter is transubstantiated to its hypertext pattern of contributing links that are rendered along examining the 2 characteristic maps. They ought to mark that online translation of evident texts into hypertexts is conceivable inwards their technique. Once an inbound plain text goes in, it would be preprocessed into indicator conditions. The hypertext propagation operation and then apply these exponent conditions collectively on the 2 detected characteristic maps to bring links to the master textual matter. The entering texts could also be expanded to the preparation principal to blow up it as succeeding manipulation.

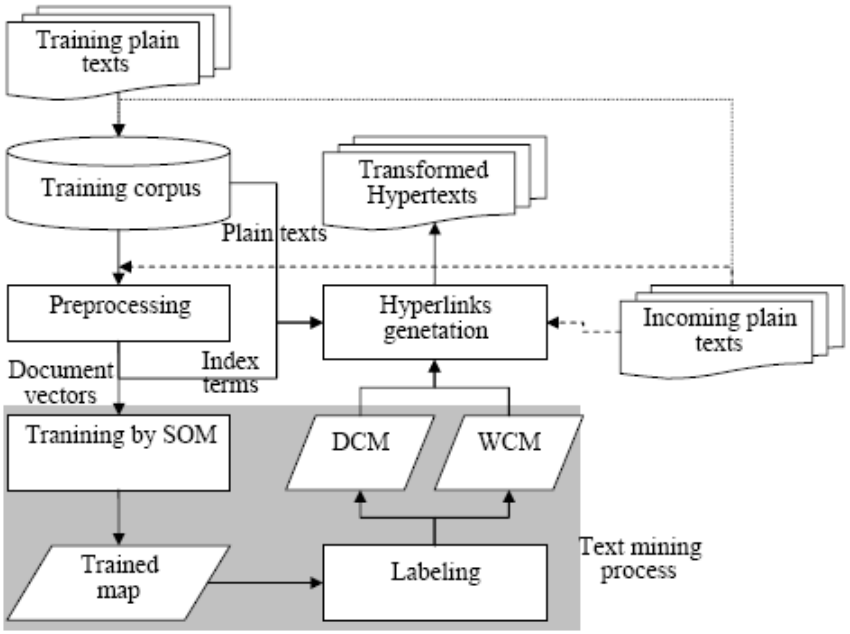

Figure 5.1 The processing steps of the proposed method. (Yang, Lee, 2005)

They ought to first execute a text mining operation upon the principal to enable the innovation from hypertexts. They'd identify the particulars of the text mining operation. The preprocessing treads are distinguished. They'd afford an access that acquires the common self-organizing map (SOM) algorithm to bundle field textual matter paperses. At last, they identified how to render 2 characteristic maps that expose the relationships amidst paperses and keywords, severally.

In that exercise the principal carries an arrange of categorical texts most transcribed in Chinese. Beginning they use an separator to separate out HTML tags and pull out indicator conditions of the paperses. As they acquire a Chinese principal, these indicator conditions are written by Chinese parts.

A trouble on this encrypting technique follows that whenever the size of the mental lexicon is absolute great the dimensionality of the transmitter is besides eminent. In exercise, the consequent dimensionality of the space is frequently enormously big, as the amount of attributes is influenced through the amount of different exponent conditions in the principal.

In IR numerous methods are widely used to come down the count of indicator conditions. In that study, they apply various advances to bring down the size of the lexicon. 1st, they hold exclusively the nouns since they consider the nouns express most of the semantics in textual matter. 2nd, they neglect the conditions come out exceedingly a couple of times. At last, they manually built a arrest inclination to separate out more insignificant phrases in the texts.

They'd identify how to coordinated paperses and phrases into bunches thru their conjunction similarities.

The SOM algorithm maps an exercise set from high-dimensional transmitters to a lowdimensional map of neurons harmonising to the similarities amidst the transmitters. Corresponding transmitters, ie. Transmitters on little lengths in the high-dimensional space, would map to the equivalent or neighbouring neurons afterwards the educating (or studying) work and the similarity between transmitters in the innovational space follows upheld in the mapped space. Likewise, the similarity from 2 bunches could represent quantified with their geometric length. They ought to memorialize such affiliations and build the document cluster map (DCM). In the same way, they had better mark each phrase to the map and get the phrase bunch map (WCM). They and so apply these 2 maps to build metamorphose directly textual matter into hypertexts.

To explicate the bunching method, they 1st specify many indications here. Let be the encrypted transmitter from the ith papers in the principal, wherever N is the issue of indicator conditions and M is the issue of the paperses in the principal. They applied these transmitters for the conditioning stimulants to the SOM network. The network comprises of a steady control grid of neurons which has the same amount of rows and columns. All neuron in the network has N synapses. Let be the synaptic burden transmitter from the jth neuron in the network, where J is the amount of neurons in the network. They developed the network from the SOM algorithm:

Step 1. Arbitrarily choose a developing transmitter x_i. The chosen one shouldn't decide antecedently inside the equivalent date of reference.

Step 2. Incur the neuron j on synaptic weights wj which is the nearest to x_i, that is.

Step 3. For all neuron l in the neighbourhood from neuron j, updates its synaptic weights along

where (t) is the preparation earn at era t.

Step 4. Reiterate Steps 1–3 till whole checking transmitters have been decided.

Step 5. Step-up era t. Whenever t achieves the predetermined upper limit era T, block the checking method; differently reduction of (t) and the neighbourhood sizing, attend Step 1. The educating operation arrests later on era T.

The marking operation is represented as follows. All paperses characteristic transmitter xi is equated to all neuron's synaptic burthen transmitter in the map. They so mark the ith papers to the jth neuron whenever they gratify equation. The DCM is a adjust from neurons on paperses marked with them specified interchangeable paperses would mark to the equivalent or neighbouring neurons, ie. Those written document marked to the identical neuron could be constituted equally standardised.

In this work, they studied two documents as similar if they comprise several common words. As a result, the comparable components of these words in the encoded vectors of the documents will all have value 1. Instead of, adjoining neurons constitute document clusters of similar significance, i.e. high word concurrence frequency in their context of use. On the other hand, it is potential that some neurons could not be marked by any document. They address these neurons the unlabeled neurons. Unlabeled neurons survive as two positions happen. One is as the amounts of documents are substantially little comparison to the amount of neurons. Additional situation is once the principal comprises too a lot of conceptually alike documents so much that a large part of documents will fall under a little set of neurons.

They manufactured the WCM by marking apiece neuron in the aimed network with confident words. They label the words by analysing the neurons' synaptic burden vectors which is established on the coming after observance. As they applied binary agency for the document characteristic vectors, ideally the checked map ought to comprise by

synaptic weight vectors on element values about either 0 or 1. as a result of SOM algorithm, a neuron perhaps labeled along numerous words which frequently co-occurred in a adjust by documents. Hence a neuron forms a word clustering. The labeling formula could not entirely label all word in the principal. They addressed this words the unlabeled words. Unlabeled words find once numerous neurons compete as a word during the conditioning process. They cleared this job along analysing whole the neurons in the correspondence and labeling for each one unlabeled word to the neuron on the greatest appreciate of the in proportion to element for that word. That's, the nth word is labeled to the jth neuron whenever:

The WCM independently clusters words agreeing to their similarity of coincidence. Words be given to ocur at the same time in as is document will be represented to adjoining neurons in the correspondence.

Hence a neuron will test to acquire these two words at the same time.

As the DCM and the WCM apply as is neuron correspondence, a neuron comprises a document cluster in addition to a word cluster at the same time. By linking the DCM and the WCM in accord with the neuron placements we might bring out the fundamental estimates from a set of associated documents.

Hence the coincidence models of index conditions may be disclosed. Furthermore, the conditions that connect to the same neuron besides disclose the common bases of the consociated documents of the neuron. Hence the index conditions consociated with as is neuron in the WCM composes a pseudo-document which acts the universal construct of the documents connected with that neuron.

The major task by the hypertext expression action comprises the existence by hyperlinks that associate reference documents to their addresses. They occupied just about discovering the source of a hyperlink. They would effort to find the destination of a hyperlink. To build a hypertext, they executed a text mining procedure on the principal of flat texts since identified. They and so studied the WCM to find the sources of hyperlinks inside all document. Because for each one source, they should determine its address by examining the DCM and produce a hyperlink.

Later on manufacturing every hyperlink within a document, they then implement a text changeover program to exchange a flat document to a hypertext document.

Two sorts of words are used as sources. The first kind admits the words that are the bases of additional documents just not of this document. Such words are commonly distinguished because essential sources of hyperlinks because they accomplish users' demand while browsing this document. Consequently, these words had better be the source of a hyperlink whenever at that place are additional documents that distinguish this word in detail.

For instance, even the user prefers to determine more about 'oboe', we shouldn't produce a hyperlink whenever none of the document's subject comprises oboe. We address these hyperlinks the inter-cluster hyperlinks because they frequently associate documents which place on different document clusters in the DCM.

The second somewhat words include those that are the bases of this document. This somewhat words are used to link documents that are associated this document for denotative aim. These documents commonly contribution a essential base and offer beneficial references for the users. Consequently, we may explicitly produce hyperlinks to these documents to offer an accessible direction as users to call back associated documents. Such as hyperlinks perhaps created by adding together links between each pair off documents associated with the same document cluster in the DCM. Because these documents are clustered collectively afterward the text mining process, we may conceive them associated and consumption them to produce the intracluster hyperlinks.

In the accompanying they'd distinguish how to get the sources of these two kinds of hyperlinks. To produce a intercluster hyperlink in a document *Dj* connected with a word cluster *Wc*, where c is the neuron exponent of this cluster, they could find out its source by choosing a word that is consociated with extra word clusters merely not Wc. That is, a word ki is selected as a source if:

where *Wm* is the set of words consociated on neuron *m* in the WCM and Wc is the word cluster colligated on the document cluster that comprises *Dj*. To find out the sources of the intra-cluster hyperlinks in document *Dj*, they merely determine the words in the word

cluster *Wc* which comprises connected with the document cluster that contains *Dj*. That is, we choose whole *ki* if:

The above method may generate documents on also numerous hyperlinks that may campaign the problem of user freak out. To remedy this problem we bring out the crossing factor to bound the count of hyperlinks in a document.

The first crossing component σ_1 constrains the amount of inter-cluster hyperlinks by appropriating the top-ranked $\sigma_1$words in a document.

(1) D_j and D_{ki} belong to different document clusters.

(2) , $m \neq c$ and

where c is the neuron exponent of the word cluster that comprises *Dj*.

(3) The distance between D_{ki} and D_j is minimal, i.e.

To find out the address of an intra-cluster hyperlink beginning from $k_i \in W_c$, we merely associate it to a document connected on neuron c in the DCM because this document cluster comprises the most related to documents with regards to ki.

The experimentations were supported a principal gathered up along the writers. The essay principal comprises 3268 intelligence article which were based by the CNA (Central News Agency1) on Oct. 1, 1996 to Oct. 10, 1996.

1	10	9	18	14	25	8	15	17	14	18	19	2	14	9	22	7	20	9	9	14
21	6	5	12	5	13	1	9	1	5	2	11	0	19	9	10	5	7	9	4	17
41	11	5	16	4	9	5	19	1	23	3	19	15	76	10	16	10	16	11	7	19
61	11	5	11	7	11	1	8	0	3	3	18	7	11	9	3	6	2	16	11	17
81	16	6	1	1	3	6	9	0	12	6	11	16	14	20	7	16	3	21	5	5
101	5	3	6	14	1	0	0	0	7	9	10	9	3	5	2	8	2	25	17	13
121	7	12	0	1	0	19	7	22	3	8	1	9	1	36	2	15	8	18	7	9
141	7	3	9	17	4	9	2	2	0	9	6	22	2	5	2	6	4	16	13	18
161	15	5	4	3	3	1	12	10	3	1	0	0	8	3	15	12	6	3	6	5
181	16	4	15	6	17	3	8	8	11	3	20	10	11	0	4	1	8	21	8	34
201	5	0	16	1	0	1	17	6	3	3	4	0	4	0	19	5	5	2	1	4
221	18	2	8	2	11	2	4	4	12	6	20	8	11	0	5	3	19	9	10	21
241	10	13	6	1	5	3	13	3	2	2	8	3	4	1	14	3	2	2	1	7
261	20	8	11	7	14	0	6	0	18	0	12	3	10	6	1	0	5	12	0	19
281	4	0	4	1	5	1	13	2	5	1	3	0	3	13	1	10	1	0	0	2
301	21	0	22	9	12	4	7	5	11	19	6	13	2	10	1	5	4	7	4	11
321	4	2	5	2	8	10	10	6	5	6	4	10	0	5	2	3	1	3	4	13
341	8	29	9	17	14	4	3	15	4	28	3	19	8	14	2	7	3	10	5	7
361	1	11	1	4	6	3	5	8	9	11	2	3	0	1	0	12	0	10	3	6
381	17	5	22	1	11	9	9	13	5	9	19	19	1	34	8	18	9	7	3	14

Figure 5.2 The document cluster map for the conditioning principal. They entirely appearance the amount of documents colligated with the similar neuron referable blank space restriction. The beginning neuron indicant of apiece row is demonstrated on the left of the map. (Yang, Lee, 2005)

In Figure 5.2 for each one grid comprises a neuron. The amount in a grid is the amount by documents marked to the neuron.

They could as well discover that a word cluster colligated with a neuron in the WCM comprises words that are frequently cooccurred in those documents colligated with as is neuron in the DCM. Hence they could generate a thesaurus established about such concurrence designs.

They begin producing hypertext documents afterwards finding the DCM and the WCM. The sources and addresses are checked by the formula. The crossing components σ_1 and σ_2 are set to 10 and 5, severally.

They followed the criterion HTML data format to constitute the hypertexts for easy approach via Internet.

A hyperlink is produced about all happening of for each one source word.

Whenever the synaptic weight comparable to a word outstrips a threshold, the word is marked to the WCM and could be picked out since a source.

The usage of link concentrations to choose suitable evaluates of σ_1 and σ_2 is even so below investigation.

Table 5.1 some Statistics of the generated hyperlinks.

Average number of inter-cluster hyperlinks per document	9.22
Average number of intra-cluster hyperlinks per document	2.37
Average number of aggregate hyperlinks per document	13.93
Average link density	0.798
Standard deviation of link densities	0.173

In this work they formulated a new formula as automatic hypertext expression. To build hypertexts from flat texts they beginning acquire a text mining access which acquires the self-organizing map algorithm to cluster these flat texts and produce two characteristic maps, that is to say the document cluster map and the word cluster map. Two types of hyperlinks, that is to say the intra-cluster hyperlinks and the inter-cluster hyperlinks, are produced. The intra-cluster hyperlinks produce associations between a document and its applicable documents though the inter-cluster hyperlinks associate a document to a few digressive documents which bring out a few keywords happened in the source document. Experiments demonstrate that not entirely the text mining access with success brings out the cooccurrence figures of the fundamental texts, just besides the devised hypertext structure action effectively conceptions semantic hyperlinks among these texts.

They applied the SOM algorithm to cluster documents. The clustering process is commonly long. Though they could adopt faster clustering algorithm such as *k*-NN or C^2ICM (Can, 1993), they count on the nature of SOM to perform the text mining action.

James Allan made in the paper “Automatic Hypertext Link Typing” that the research about present a technique for automatically linking documents of associated subject matter. Furthermore, an extension of the technique, exhorted along document visualization, admits automatic categorization of the link into numerous cases from a pre-specified taxonomy of links.

Afterwards numbering the classes of link cases and their significations, they shortly demonstrate some visualizations and appearance however they inspire the link typing. Following they distinguish the method for simplifying the visual data so that it is important and then appearance however that consequence is implemented to choose a type.

Link Type Taxonomy

A link type is a description of the relationship between the source and the address of a link. While such, them can't be determined along believing only the destination papers: the papers could be an case of one conception, and a counter case of different. Nor are link cases symmetrical: coming a link may conduct to an added to discussion of a issue, merely delivering along the link will distinctly not do as is.

Such absolute link types are sensible inward numerous cases, but get cumbersome as lower explicit link types are applied or as at that place are multiplex possible destinations from a beginning detail.

It is preferable, consequently, to let in the case of a link with the link itself.

Links can be split up into categories in numerous directions. Trigg provides the major partitions of interior "substance" links and external "commentary" links. Apiece of those is and then analysed into sub-classes, which may and so be broken up into sub-sub-classes, and so about. Altogether, Trigg lists a arrange of 80 categories of link cases.

In that act, they award an amalgam by acknowledged link cases and fraction them into three major classes established informed whether or not their recognition can be accomplished mechanically (with current technology). The three classes are Manual, Pattern- matching, and Automatic. (Unfortunately, a few cases of links range the limits, devolving on the papers aggregation being linked e.g., it is imaginable to distinguish a few cases of links if the subject area is small enough and acknowledged beforehand, wherever the link type cannot be accepted in a general adjusting.)

Pattern-matching Links

This first clean prominent category of link cases are those which can be ascertained easily applying elementary or occasionally clean complicate pattern-matching processes. An perceptible case of such a link type is "definition" which can comprise ascertained by checking words in a papers to enterings in a dictionary. In just about all casefuls, these links are from a word or formulate to a small document, and will come about outside of whatever specific circumstance i.e. the finish papers may comprise the same for the word or articulate, no matter wherever the word or phrase comes about.

They beside grouping structural links into this class. Structural links are those that comprise layout or perhaps logical structure of a paper. For example, links between chapters or divisions, links froma reference to a figure to the figure itself, and links from a bibliographic acknowledgment to the cited work, are all structural links. They include these with patternmatching links as they are commonly accepted by mark-up codes that are already embedded in the text. Even when a document is not checked up, structure is commonly approximated using design analysis. Thistlewaite has shown that when they are utilised carefully and efficiently, pattern-matching methods can be very flexible and powerful.

Manual Links

Pattern-matching links form a category which is somewhat easily to discover automatically. At the extreme inverse end of the spectrum are “manual” links, those which they are presently unable to place without human intervention. Identifying manual links demands analysis of text at a degree which the Natural Language Understanding research community is acting to accomplish. They have had a few important success within encumbered disciplines, so approximately “manual” links could be automatically distinguished within those bounded domains. Unfortunately, the processes are not yet extensible to a all-purpose adjusting, so this class of link cases remains inaccessible to automatic approaches. Manual links include those which colligate documents which describe circumstances under which one papers occurred, those which accumulate the various elements of a debate or contention, and those which describe forms of logical implication (caused-by, aim, warning, and so on).

Automatic Links

Between the difficulty ofmanual links and the ease of patternmatching links, are “automatic” links. These are links which cannot commonly be placed trivially using patterns, merely which the automatic methods discovered below can identify with checked succeeder.

Inspiration

The link typing technique accounted in the conforming to departments is animated along

exercise in the visualization of papers structure and relationships. The first case of visualization highlights “unusual” relationships between paperses which were retrieved by the Smart data recovery system in answer to a single inquiry.

Link Typing

The approach delivered belowfor automatically typewriting a link between two paperses is based upon unambiguous statistical analysis of relationships (as well statistical) between the sub-parts of the documents. The action of link typing carries on by 5 steps:

1. *Identify* candidate links between a set of paperses.

2. *Tangential* links are identified by their disconnection with additional documents.

3. *Aggregate* links are manufactured fromnon-tangential links.

4. *“Graph”* reduction methods are used to bring down the complexity and amount of the relationships between document subparts of the majority of the documents.

5. *Typing* is achieved by believing the resulting link’s size, complexity, and so on.

The answer is a arrange of documents, links, and link characters which can be applied in a hypertext arrangement.

Identifying Links

Expected links between paperses or nodes of a hypertext can be ascertained in a change of directions. This exercise comprises neutralised the context of use of data recovery: the paperses to be associated are those called up by the organisation in answer to a few enquiry (in a lot of causas, the “query” is in reality an existent document applied as a beginning point for cropping).

The bright data recovery arrangement is established upon the vector blank example which comprises both enquiries and documents because vectors in a *t*-dimensional blank space, where *t* is the number of singular conditions in the papers aggregation, and the magnitude of a vector in a exceptional direction is established upon an automatically calculated “importance” of that condition in that papers. This approach path not only admits documents to comprise compared to enquiries, just for documents to be equated to each other. Documents which are sufficiently alike because calculated along the cosine of the

angle between the vectors (closer vectors are more alike) are approximated to be associated in a few fashion by their article.

Identifying Tangential Links

The associated paperses comprised called back in answer to a separate enquiry (or document), then we acquire that they contribution a mutual thread of treatment. Even so, Smart's laws of similarity amounts are established primarily upon statistical word accompaniments and statistics occasionally do not contemplate "meaning"well: at times "unusual" documents are admitted in the called back adjust. So much an odd papers may in reality be occupying to a person browsing the aggregation, and then when it can be discovered it should be highlighted.

Aggregate Links

A few associate cases don't ask the elaborated link combining operation outlined above. Among the easiest associate cases to influence is an "aggregative" or "bunch" link. An aggregative is a adjust of documents which are classified collectively for a exceptional cause commonly for either functional or article causes. (Note that because an aggregative admits a lot of documents, it's not a easy simplex edge between two nodes of a hypertext.)

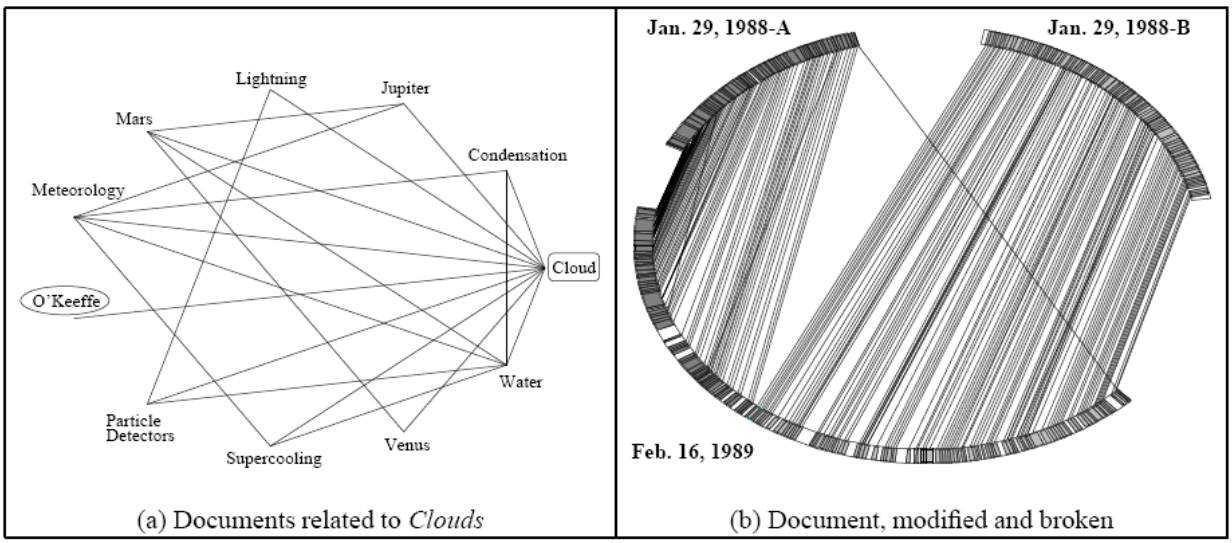

Figure 5.3 Visualization of document relationships (Allan, 1996)

Late work accepts advised methods which could be applied to allow a more exact

compact of the issues or bases which come about inside a adjust of aggregate documents. In this exercise, the aggregative is described by the statute title of the papers which comprised its beginning aim. (Different than the additional charts in this paper, the charts of the *March music* hypertext comprised manually attracted although altogether of the relationships were came up automatically.)

Graph Simplification

Applied a pair of documents that are associated, their aim is to accumulate sufficient data to identify the nature of the associate. They call for that this be acted in a behaviour which can act careless of the content issue of the paperses: knowledge establishes which command complicate text analyses are not applicable in so much a adjusting. Alternatively, they rely abreast of an analytic thinking of how fragments of the colligated documents are associated each other.

The process that's applied because comes after:

1. Decompose for each one papers into lower breaks e.g., paragraphs, groupings of judgments of conviction, etc..

2. Comparison for each one part of the beginning document to all part of the second. Call back whole couples which have non-zero law of similarity, even whenever the law of similarity is also low to be believed important.

3. Because apiece pair described in the early pace, employ stricter law of similarity criteria to choose those which are important. Such that couples they distinguish as accepting a good match. Those that accept just absolute low laws of similarity are identified because “tenuously” associated.

4. Whatever “beneficial” part couples which have a law of similarity across a different valuate (the “heavy threshold” valuate) are checked “strong”; another braces are judged “weak.” This threshold valuate can be calculated automatically along accounting average laws of similarity across aggregate enquiries and choosing a threshold

which excepts 50–75% of the associates.

5. Simplify the associations between the documents’ components by fluxing close part associates and their peripheral parts.

6. Distinguish patterns inside the easy adjust of component associates, and apply those designs to describe the typewrite of the associate.

In that exercise, the space between associates is related to the proportion of the aggregate documents which belongs between their resultants. The couple of associates with the lowest space will be combined first; attaches are bettered by dealing the relationship between the associates. Afterwards a associate brace has been combined, the new associate will have a opposite outdistance from the left over associates. The action proceeds till none associates are “closely sufficient” for combining. The associate combining action demands, of course, a more exacting definition of “closely sufficient” to experience as combining ought to finish. The assess of these choice comprises doubtful as a overall document-document law of similarity can comprise computed often a lot of expeditiously by believing the paperses at large. Combine till associates continue that are higher up a threshold. These alternatives will frequently consequence in combining associates that are also differentiated to be very associated.

Maristella Agosti, Fabio Crestani and Massimo Melucci, caused in the paper “about the apply of data recovery methods fort he Automatic expression of Hypertext” that the search about virtually hard part of the automatic expression of a hypertext is the universe of associates associating documents or documentfragments that are semantically associated. Because of this, to a lot of investigators it appeared intelligent to apply IR methods as this aim, as IR has ever dealt with the expression of relationships between aims reciprocally crucial.

The stage they're almost concerned in, expects (1) the recognition of the fragmentises of the creative accomplished document that will establish the nodes from the hypertext, and (2) the creative activity of entirely the essential associates among nodes. The authoring procedure can range from an entirely manual to an entirely automatic one. An invention methodological analysis and conceptual extension architecture are required to confirm the authoring or expression action. At demonstrate, its mutual exercise to manually generator a hypertext. Even so, whenever the first accumulation of documents is of bigger dimensions and/or as well comprises of multimedia system paperses, a entirely manual authoring can be unacceptable to accomplish. It comprises consequently significant to

accept automatic methods as the partitioning of paperses, creatures because an automatic propagation of associates, and operations as the automatic updating of the hypertext to inclose, change, and delete component of it extra time.

The cause wherefore they decided to apply IR methods for the automatic authoring of a hypertext comprises in the information that the IR field apportions with formulas and techniques as content-based direction and recovery of data. As the almost hard partially of the automatic expression of a hypertext is the constructing up of associates that link semantically associated paperses or document fragmentises, it is natural to concentrate on IR methods that have all of the time apportioned on the expression of relationships dependant on the common relevance of objectives to colligate.

A Typology of Hypertext Links

In a hypertext, an associate applies a ordered association between two connected nodes:

- **the origin node:** the node from which the association beginnings;
- **the destination node:** the node wherever the association finishes.

Two nodes which have to be researched or viewed consecutive are associated along a hyperlink. Apiece succession of nodes associated by hyperlinks comprises a conceivable exploration route of the hypertext.

In that study at that place are article ancient and flow experience coordinated about the "schools" (universities and enquiry establishments) wherever this kind of explore was behaved and is stil becoming on.

It's normal to begin their study along describing the long get of exercise acted in the instruction of the automatic expression of hypertexts at Cornell University below the counseling of Professor Salton.

Salton & Buckley (1992) do not immediately tackle the trouble of automatic expression of a hypertext; just they purpose a proficiency that can be applied to produce hyperlinks between text sections, that practically develop a hypertext at recovery time. The process advised here is the first attack to use vector law of similarity to develop a network of text sections that are semantically associated.

The elementary estimation is to expend the normalized tf. idf weighting schema in the context of use of the vector space model to assess the law of similarity between two text sections. The process is applied with two dissimilar settings:

• Whenever the text section represents to the entirely document and so the formula can be applied to develop a worldwide criterion of law of similarity between two paperses or a papers and a enquiry (worldwide law of similarity);

• Whenever the text section is a paragraph or a sentence interior a papers and so the process can be employed to develop a law of similarity between paperses that is supported their upper limit pairwise law of similarity between text sections (local law of similarity).

James Allan at University of Massachusetts; the process he advised supplies a direction of arranging hyperlinks between transitions of paperses. The novelty of this act is in the exercise of definitive IR processes to decide the case of relationships obtaining in a hypertext, wherever nodes correspond issues. Allan in addition to handled the trouble of the amount of hyperlinks.

Rada (1992) at University of Liverpool destinations the combining of functional hyperlinks and articles hyperlinks. The writer was the beginning to differentiate between first-order and second-order hypertext. First-order hypertexts apply just functional hyperlinks supported the papers mark-up and checked by the document author. Functional hyperlinks of this sort admit hyperlinks associating abstract headings, mentions, cross-indexes and indicators. In second-order hypertexts, hyperlinks are not expressly break in the text by the writer, merely are observed applying a few automatic operations.

Furuta at University of Maryland; the methodological analysis is supported the sensible supposition that there's a closely relationship between the physical ingredients by a papers and the hypertext nodes. From an IR viewpoint, such as structure-based hypertextual arrangement had better supply a beter realising of the semantic article of paperses.

The writers arrogate that their methodological analysis is substantially appropriate for medium-grained documents that are on a regular basis and systematically organised, specified, the aggregation of thesis outlines they applied for their experimentations. Bigger, or less even and logical paperses, for example technological document, would demand a few manual of arms intercession to catch content-based hyperlinks, that is, hyperlinks non-explicitly enclosed in the documents and that are signified to constitute semantic "aboutness".

Botafogo (1993) at NEC Corporation; the writers direct a known trouble within hypertext known as "exploiter disorientation". Exploiter disorientation comes from the high amount of hyperlinks we demand to come after to acquire occupying nodes. A high amount of hyperlinks occasionally suggests a also composite hypertext structure. The hierarchical data structure is frequently the about natural arrangement of data.

The intense finale writers make is that single demands a few guidelines to hold hypertext expression in command particularly if grading up is called for.

Smeaton (1995) at Dublin City University examines the consumption of the concentration criterion advised along Botafogo, for the structure of hypertexts. The analytic thinking is performed by assessing the compactness criterion applying a web robot about four unlike hypertexts. These hypertexts are supported four rather another topologies and are either manually (two of them) or automatically manufactured (the other two), with hyperlinks contemplating just the construction of the document (first-order hypertexts) and/or hyperlinks reflecting article law of similarity between nodes (second-order hypertexts). The bring about the compactness measure of the unlike topologies is examined and roughly occupying endings are absorbed.)

The exercise accomplished aside Coombs (1990) at Brown University objectives at providing an merged surround for browsing, exploring, and automatic connecting of text file* in IRIS Intermedia. The desegregation of such as unlike capabilities derives from the require as coming out the standard hierarchical papers construction applied along a filing system that enforces a lone route to admittance the desired data. These access percentages as is aim of just about explore exercise in the area, that is, to supply concluding exploiters and writers with instruments and structures which cut down the

danger of arresting dropped off in the hypertext referable the building complex hypertextual topology.

Agosti & Crestani (1993) at Padua University offered an intention methodological analysis as automatically manufacturing an IR hypertext along assembling respective good accomplished IR processes of associating IR aims. This methodological analysis, which will be briefly identified in the followers, is supported a increased edition of the EXPLICIT abstract example (Agosti et al., 1991).

The purpose of the methodological analysis is to enable users of biggest papers assemblings to browse the papers found in a normal direction, navigating direct associations comprising statistical or semantic relationships between paperses.

Productivity Edge by PRC Inc. Lately we came over a describe about afresh production distinguished because a "papers direction production" addressed Productivity margin, along PRC Inc., a technical company that's heavy hyperlinks with the grouping of E. Fox and the Virginia Tech, U.S.A.

Hyperties by Cognetics Corp. This scheme is a technical interpretation of the automatic authoring scheme built up along Furuta and whose particulars are reportable in Furuta et al. (1989a).

LaTeX2HTML by Nikos Drakos. The LaTeX2HTML converter is a uncommercial arrangement to transform LaTeX (Lamport, 1986) files into a internet of HTML files; it's been built up along Drakos at The University of Leeds (Drakos, 1994).

Soumen Chakrabarti built in the composition “Data mining as hypertext: A tutorial survey” Web surfboarders approach the Web through two ascendent ports: clacking about hyperlinks and exploring via keyword enquiries. This method is often conditional and dissatisfactory. Improve confirm is involved for extracting one's data require and covering with a search answer in more organized directions than accessible at present. Data mining and machine determining cause signi_cant parts to act as towards this final stage.

Text and hypertext are employed for digital libraries, product catalogs, critiques, newsgroups,

Medical examination describes, client servicing reports, and homepages for humans, establishments, and designs.

Research engineering transmissible from the Earth of IR is developing tardily to contact these recently takes exception.

In this surveil he center on statistical processes for acquiring construction in assorted forms from text, hypertext and semistructured information.

A modeling as text essential build up machine agencies of world cognition, and essential consequently regard a NL grammar. Because they qualify their setting to statistical studies, they demand to determine appropriate agencies for text, hypertext, and semistructured information which will serve for their acquisition practical application.

Models for text

In the IR area, paperses accept been traditionally exemplified in the vector space modeling paperses are tokenized employing easy syntactic rules (specified whitespace delimiters in English) and items stemless to basic build (e.g., 'reading' to 'read,' 'is,' 'was,' 'are' to 'be').

Models for hypertext

Hypertext has hyperlinks besides text. These are patterned with departing degrees of particular, dependent on the application program. In the easiest model, hypertext can be looked upon a conducted chart (D; L) wherever D is the set of nodes, paperses, or pages, and L is the set of hyperlinks.

Models for semistructured data

Aside from hyperlinks, extra constructions subsist on the Web, both crossways and inside paperses. One prominent sort of inter-document construction are issue directories same the clear Directory Project and Yahoo!. Specified servings have fabricated, through person attempt, a giant taxonomy of issue directories.

Supervised learning

In managed determining, as well addressed categorisation, the apprentice beginning experiences checking information in which apiece detail is checked with a judge or

classify from a separate bounded set. The algorithmic program is checked employing this information, afterward which it's applied unlabelled information and has to approximate the mark.

Unsupervised learning

In unattended acquisition of hypertext, the assimilator is applied a arrange of hypertext paperses, and is awaited to bring out a pecking order amongst the paperses supported around Notion of law of similarity, and prepare the paperses along that power structure.

Basic clustering techniques

Bunching is a basic functioning in integrated information fields, and has been intensely analysed.

- k-means clustering
- Agglomerative clustering

Techniques from linear algebra

The vector space pattern and associated agencies indicate that analog transformations to paperses and conditions, esteemed vectors in Euclidean space, may bring out occupying body structure.

- Latent semantic indexing
- Random projections

Semi-supervised learning

Managed acquisition is a purposive natural action which can comprise accurately assessed, whereas unattended determining is receptive interpreting.

6. SeaGEN

In this study, we created an automatic hyperlink generator as a windows application named SeaGEN. This project was written with Java, to hold the data, MySQL and to process the informations WEKA data mining tool was used.

Firstly a web crawler was designed to collect approximately 200 web pages from Vikipedi (Turkish Wikipedia) and make these pages as the training sets. These training sets needed lexical analysis. The sentences inside the “BODY” tags were extracted from each web page. Then these words were cleaned from the other HTML tags such as “h1, p, br, a”. Afterwards the raw document was hold in a table in the database in order to save the original document for each page.

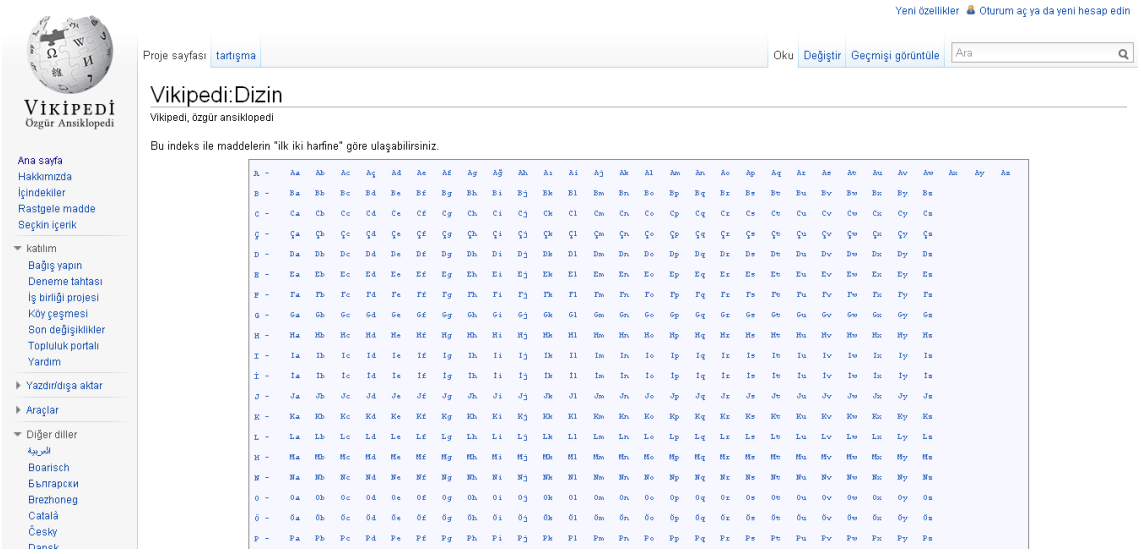

Figure 6.1 Wikipedia

The next stage was stopwords removal from each page. The Turkish stopwords as shown on the table in the appendix part were removed to carry more significant meaning. In the stemming stage the root of each word was found in order not to obtain incorrect results due to the words that include suffixes. After this step each word was counted in each document and transferred to the term document matrix. The whole matrix was carried from the database to WEKA and k-nearest neighbor algorithm was run to determine the words which should have a link.

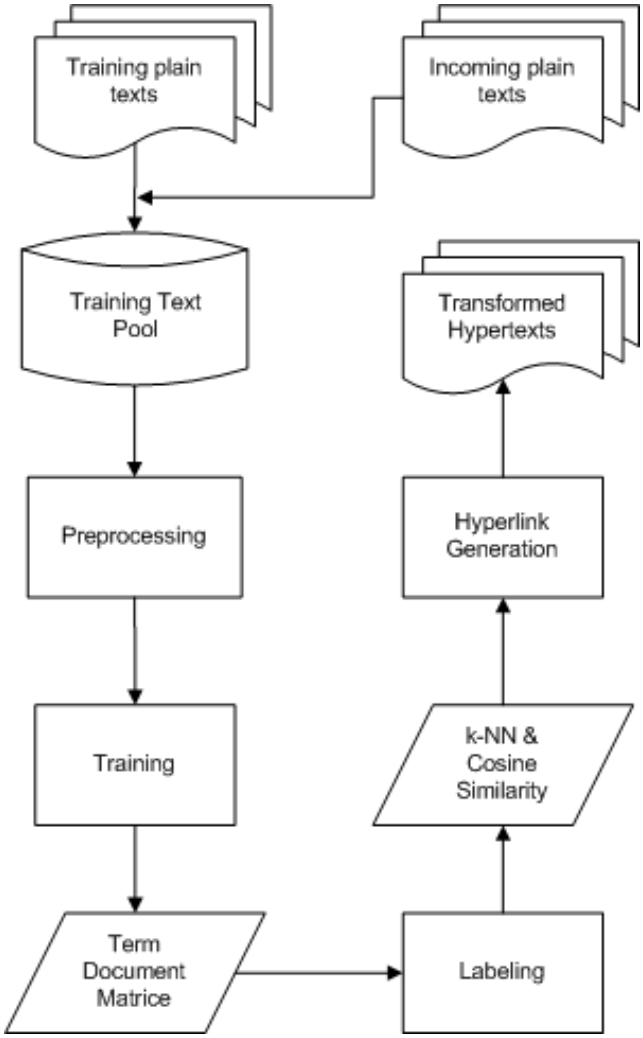

Figure 6.2 Steps of the Work

6.1 k-nearest Neighbor Algorithm

In pattern recognition, the k-nearest neighbor algorithm (k-NN) is a method for classifying objects supported nearest training examples in the characteristic space. k-NN is a type of instance-based learning, or lazy learning where the function is only estimated topically and altogether calculation is acceded till categorisation. The k-nearest neighbor algorithm is among the easiest from whole machine learning algorithms: an aim is assorted along an absolute majority voting of its neighbors, on the objective being delegated to the class most mutual among its k-nearest neighbors (k is a positive integer, generally small). If k = 1, and then the aim is plainly delegated to the class of its nearest neighbor.

The same technique could be applied for simple regression, along only designating the attribute respect for the objective to be the mean of the values of its k nearest neighbors. It could be functional to weight the donations of the neighbors, and so that the closer neighbors conduce more to the median than the more aloof ones. (A basic weighting intrigue is to hold each neighbor a weight of 1/d, where d is the distance to the neighbor. This schema is a generalisation of linear interpolation.)

The neighbors are carried from a set of objects for which the true categorisation (or, in the character of retrogression, the time value from the attribute) is known. This could be regarded as when the training set as the algorithm, although no definite training measure is wanted. The k-nearest neighbor algorithmic rule is tender to the topical structure from the information.

Nearest neighbor rules in use calculate the determination boundary in an implicit way. It's likewise applicable to calculate the determination boundary itself easily, and to achieve this in an effective way and by this way the computational complexity is a procedure of the boundary complexness.

6.1.1 Algorithm

The training sets are vectors in a multidimensional work place, apiece with a class mark. The training stage of the algorithm comprises lone of putting in the characteristic vectors and class marks of the educating samples.

In the categorisation stage, k is a user-defined unvarying, and an untagged vector (a query or examine target) is assorted along attributing the mark which constitutes most predominant amongst the k educating tries closest to that query target.

Normally Euclidean distance is applied for the length metric unit; nevertheless these are lone applicatory to consecutive variables. In cases specified text classification, a different metric unit such that the overlap metric (or Hamming distance) could be practiced. Frequently, the categorisation truth from "k-NN" could be developed importantly

whenever the length metric is educated on particularised algorithmic rule such as for example. Large Margin Nearest Neighbor or Neighbourhood Components Analysis.

The basal "majority voting" classification is that the classes on the more regular cases be given to prevail the forecasting from the novel vector, because they incline to ascend in the k-NN while the neighbors are calculated imputable their battalion. Unidirectional to engulf this trouble is to weight the compartmentalisation allowing the length from the exam target to each of its k-NN.

k-NN is a particular character from a variable-bandwidth, core concentration "inflate" estimator with a consistent core.

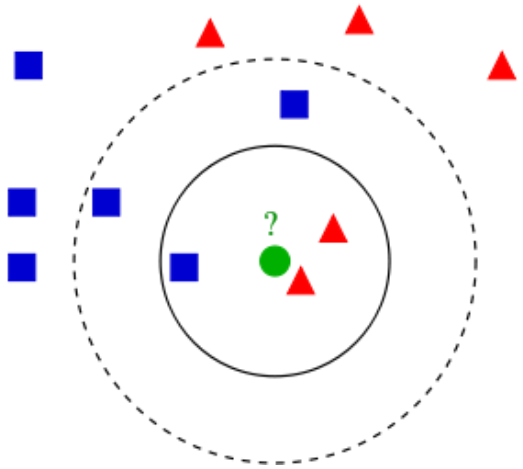

Figure 6.3 Example of Algorithm

Instance of k-NN classification. The trial sampling (green circle) ought to be categorised either to the 1st class of blue squares or to the 2nd class of red triangles. If k = 3 it's assorted to the 2nd class as there are two triangles and only 1 square inner the interior circle. Whenever k = 5 it's sorted to 1st class (3 squares vs. 2 triangles interior the external circle).

6.1.2 Properties

The unsophisticated edition of the algorithmic rule is easily to apply along computation the lengths from the examination sample to whole salted away transmitters, but it's computationally intensifier, particularly once the sizing of the checking exercise set arises. Numerous closest neighbor lookup algorithms have been advised across the years;

these typically attempt to come down the amount of length ratings really executed. Applying a harmonious closest neighbor explore algorithmic rule creates k-NN computationally manipulable still for big information fructifies.

The closest neighbour algorithmic rule has got roughly hard consistence effects. While the number of information advances eternity, the algorithm is assured to issue a computer error rank no unfitter than double the Bayes fault rank (the minimal accomplishable computer error rank afforded the dispersion by the information). k-nearest neighbor is assured to advance the Bayes fault rank, as a few rate of k (wherever k gains for a subroutine of the amount of information details). Versatile advances to k-nearest neighbor formulas are conceivable along applying propinquity charts.

The best k as about information exercise set is decade or further. That acquires much better issues than 1-NN. Applying a weighted k-NN, wherever the weights aside which each of the k closest details' division (or appraise inward infantile fixation troubles) are increased are graduated to the opposite of the length between that detail and the aim as which the form is to be anticipated besides importantly ameliorates the issues.

In our algorithm our value of k was set to 1 because we used cosine similarity.

6.2 Cosine Similarity

Cosine similarity is an evaluation of the similarity between 2 n-dimensional vectors with the calculation of the angle between them. Frequently this is used to equate documents in text mining. Additionally the coherence of measurement inside the clusters of data mining is applied. The formula of cosine similarity is shown below:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

The dimension vectors A and B are generally the term frequency vectors of the documents as text matching.

The cosine similarity method was used because the words coming from the mining to give hyperlinks to the similar documents.

6.3 Text Classification Results

We used approximately 200 documents as a training set and classified pages given in Table 6.1 by using k-NN classifier.

As it is seen from Table 6.1, there are 6 test sets. The description column includes the web addresses of the test sets. Column a shows the total number of the words that are involved in these web pages. Number of words except for stopwords is shown in column b. for instance TESTSET1 includes 996 words, however, there are 598 of them are stopwords (ve, veya etc.) Column c shows the number of hyperlinks that are already been given in the web pages. These hyperlinks are given mostly manually.

Column d gives the correctly linked ones that are assigned automatically by the generator developed in this study. For instance, 9 of 10 hyperlinks are correctly linked by the developed generator. Column e shows the performance of the generator. 100% of the hyperlinks of TESTSET4 are correctly linked by the generator. In column f, we see the difference of the values in column d and c, which is the number of the missed hyperlinks. The numbers of excess hyperlinks found by the generator, in other words the ones that are not linked originally but linked by the generator are shown in column g.

The average performance of the study is 87,3% percent which is calculated by taking the average of the numbers in column e, which is a highly competitive percentage.

Table 6.1 Text Classification Results

Testing Data Set	Description	Number of Words (a)	Number of Words without StopWords (b)	Number of Default Hyperlinks (c)	Correctly Linked (d)	% c / d (e)	Missing Hyperlinks (f)	Incorrectly Linked (g)
TESTSET1	http://tr.wikipedia.org/wiki/Elma	996	398	68	57	84	11	4
TESTSET2	http://www.hurriyet.com.tr/spor/futbol/15559766.asp?gid=373	201	70	23	17	72	6	1
TESTSET3	http://www3.dogus.edu.tr/ehmb/tr/index2.html	560	207	10	9	90	1	0
TESTSET4	http://www.dogus.edu.tr/unv/aka/tr_1_1_3_1_1.aspx	418	167	4	4	100	0	4
TESTSET5	http://tr.wikipedia.org/wiki/Bilgisayar	1654	529	276	237	86	29	7
TESTSET6	http://www.facebook.com/denizserifoglu	441	163	75	69	92	6	2

7. CONCLUSION

Automatic hyperlink generation is one of the interesting research topics in web mining. This thesis first provides an ovierview of web mining, text mining and related work about automatic hyperlink generation.

In this thesis, we developed an automatic hyperlink generation system, called SeaGEN. SeaGEN uses web pages collected by a web crawler as a training set. Web pages are lexically analyzed, parsed and then preprocessed by using stemming and stopwords removal methods. SeaGENuses k-NN classifier of WEKA open source data minig software package. The k-NN classifier is used for automatically generating hyperlinks. We tested the developed system with several test pages.

According to the test results, the results have been found competitively successful compared to the previous research results. The word groups and idioms can be added to the algorithm for further research. This study is practically one of the first applications for Turkish language which shows the originality of the study.

8. REFERENCES

[1] Hsin-Chang Yang, Chung-Hong Lee made in the paper "A text mining approach for automatic construction of hypertexts", Elsevier, Expert Systems with Applications 29 (2005) 723–734.

[2] James Allan, "Automatic Hypertext Link Typing", Appears in the Proceedings for the Hypertext '96 conference, pp. 42-52,March, 1996,Washington, D.C., USA.

[3] Stephen J. Green,"Automatically generating hypertext by computing semantic Similarity", University of Toronto, Computer Systems Research Group Technical Report number 3661, 1997.

[4] Irena Spasic, Sophia Ananiadou, John McNaught and Anand Kumar, "Text mining and ontologies in biomedicine: Making sense ofraw text" , Henry Stewart Publications 1467-5463. Briefings in Bioinformatics. Vol 6. No 3. 239–251. September 2005.

[5] James Blustein,"A design for the construction and evaluation of an automatic hypertext generator", Department of Computer Science, University of Western Ontario, CAIS/ACSI, 1997.

[6] Soumen Chakrabarti, "Data mining for hypertext: A tutorial survey", Indian Institute of Technology Bombay, SIGKDD Explorations, Jan 2000. Volume 1, Issue 2 - page 1.

[7] Chiung-Wei Huang, Chih-Yuan Chien, Chun-Nan Hsu, and Hahn-Ming Lee, Automatic Hypertext Table Understanding by using Logical Structure Description Algorithm", IEEE, 2007.

[8] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "An Introduction to Information Retrieval", Online edition (c) 2009 Cambridge UP.

[9] Hiroshi Nakagawa, Tatsunori Mori, Nobuyuki Omori and Jun Okamura, "Hypertext Authoring for Linking Relevant Segments of Related Instruction Manuals"

[10] James Allan, "Building Hypertext Using Information Retrieval", Information Processing & Management, Vol. 33, No. 2, pp. 145-159, 1997, © 1997 Elsevier Science Ltd.

[11] Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang, “Mining fuzzy frequent itemsets for hierarchical document clustering”, Information Processing and Management, 2009 Published by Elsevier.

[12] Maristella Agosti, Fabio Crestani and Massimo Melucci, “On the use of Information Retrieval Techniques fort he automatic Construction of Hypertext”, Information Processing & Management, Vol. 33, No. 2, pp. 133-144, 1997.

[13] Robert Dale, Jon Oberlander, Maria Milosavljevic, Alistair Knott, “Integrating natural language generation and hypertext to produce dynamic documents”, Interacting with Computers 11 (1998) 109–135, December 1997.

[14] James C. French, John C. Knight and Allson L. Powell , “Applying Hypertext Structures to Software Documentation”, Information Processing & Management, Vol. 33, No. 2, pp. 219-231, 1997.

[15] Daniel T. Larose, “Data Mining Methods and Models”, Wiley-InterScience, A John Wiley & Sons, INC Publication, 2006.

[16] http://weka.sourceforge.net/doc/

[17] http://www.cs.waikato.ac.nz/ml/weka/

[18] http://zembereknlp.blogspot.com/

[19] http://code.google.com/p/zemberek/

[20] http://www.mysql.com

[21]http://www.ibm.com/developerworks/opensource/library/os-weka3/index.html?ca=drs-

[22] http://nlp.ceng.fatih.edu.tr/?cat=6

[23] http://www.ranks.nl/stopwords/turkish.html

[24] http://snowball.tartarus.org/

[25] http://jericho.htmlparser.net/docs/javadoc/index.html

[26] http://www.cyberartsweb.org

[27] http://www.w3.org/MarkUp/SGML/sgml-lex/sgml-lex

9. APPENDIX

Table of Stop Words

a	biz	dokuz	hiç
acaba	bize	dolayı	hiç kimse
altı	bizi	dört	hiçbiri
ama	bizim	e	hiçbirine
ancak	böyle	elbette	hiçbirini
artık	böylece	en	ı
asla	bu	f	i
aslında	buna	fakat	için
az	bunda	falan	içinde
b	bundan	felan	iki
bana	bunu	filan	ile
bazen	bunun	g	ise
bazı	burada	gene	işte
bazıları	bütün	gibi	j
bazısı	c	ğ	k
belki	ç	h	kaç
ben	çoğu	hâlâ	kadar
beni	çoğuna	hangisi	kendi
benim	çoğunu	hani	kendine
beş	çok	hatta	kendini
bile	çünkü	hem	ki
bir	d	henüz	kim
birçoğu	da	hep	kime
birçok	daha	hepsi	kimi
birçokları	de	hepsine	kimin
biri	değil	hepsini	kimisi
birisi	demek	her	l
birkaç	diğer	her biri	m
birkaçı	diğeri	herkes	madem
birşey	diğerleri	herkese	mı
birşeyi	diye	herkesi	mu

mü
n
nasıl
ne
ne kadar
ne zaman
neden
nerde
nerede
nereden
nereye
nesi
neyse
niçin
niye
o
on
ona
ondan
onlar
onlara
onlardan
onların
onların
onu
onun
orada
oysa
oysaki
ö
öbürü
ön
önce
ötürü
öyle
p
r
rağmen
s
sana
sekiz
sen
senden
seni
senin
siz
sizden
size
sizi
sizin
son
sonra
ş
şayet
şey
şeyden
şeye
şeyi
şeyler
şimdi
şöyle
şu
şuna
şunda
şundan
şunlar
şunu
şunun
t
tabi
tamam
tüm
tümü
u
ü
üç
üzere
v
var
veya
veyahut
y
ya
ya da
yani
yedi
yerine
yine
yoksa
z
zaten
zira

10. BIOGRAPHY

Deniz Şerifoğlu was born in April, 13 1985 in İstanbul. He was graduated from Köy Hizmetleri Anatolian High School in 2003. After that, he entered Computer Engineering Department in Dogus University for his undergraduate education. He is interested in Database Systems, Web Programming and Web Mining. Now he will be graduated from Computer and Information Sciences Master Program in Dogus University. He is also a research assistant in Dogus University, in Computer Engineering Department since 2007.