Integrated Visual Analytics Tool for Heterogeneous Text Data

Jihyoun Park*

Hong Kong Applied Science and Technology Research Institute

ABSTRACT

Our self-developed java-based visual analytic tool reads a variety of different text data sources and extracts important keywords, relations and events from them using ontology and natural language processing methods. Finally it provides an integrated and interactive search interface to users to facilitate their effective and efficient investigation for the large and complex data set.

Keywords: VAST 2014, mini challenge 1, visual Analytics, ontology modeling, natural language processing.

Index Terms: [Visualization]: Visualization application domains—Visual analytics

1 INTRODUCTION

Visual analytics is defined as the science of analytical reasoning helped by software using interactive visual interfaces [1]. As a combination of visualization and automatic analysis, visual analytics can be more effective and efficient than each solution alone, if the data set provided to solve a problem contains many different types of data, thus requires a different level of human and machine involvement to properly deal with them [2].

The mini-challenge 1 asked us to investigate the background of two major organizations related with the recent kidnapping incident in Kronos, POK and GAStech, to find out facts, which may be related to or affected this incident. We were also required to summarize events for two days when 14 GAStech employees were found missing by reading a variety of data sources. The challenge provided 6 different types of text data to tackle this issue. Firstly, we created software to automate the analysis and aggregation of data in order to help organize and extract information from the massive data. Besides, we added a user interface to manually verify and modify the automatically analyzed results so that users could check the basis of their decision making. Then the final aggregator and query interface supported human investigators to find out the evidence to answer questions.

The rest of the paper is organized as follows. In section 2, the text data analysis methodologies adopted in this paper are explained. Section 3 introduces the visual analytics tool that we created to solve the mini-challenge 1. Section 4 concludes with the summary and findings.

2 DATA ANALYSIS

Hogenboom et al. [3] classified methods to extract events from text roughly into two categories: data-driven and knowledgedriven. The latter has again two sub-methods: lexico-syntactic patterns and lexico-semantic patterns according to the characteristics of text. As our data sources have different formats

IEEE Symposium on Visual Analytics Science and Technology 2014 November 9-14, Paris, France 978-1-4799-6227-3/14/\$31.00 ©2014 IEEE and types, we employed the hybrid method. For example, we used the ontology modelling, which is a lexico-semantic pattern, to build relations of organizations and people, and lexico-syntacticbased keyword extraction methods and data-driven methods to summarize news with a set of important keywords.

Before applying those methods, we analyzed the types of given data (Figure 1). Structured data such as GAStech employment records and email headers of internal GAStech employees can be handled automatically to analyze the contents. We classified countries' fact sheets and resumes as semi-structured data, because they had clearly divided sections but some contents were written in a free-style text format. After our tool roughly created the ontology model based on the title of each section, users were required to review the structure manually. We also had unstructured data such as POK history reports. In this case, we decided to input the ontology model of the data manually for more accurate analysis. Lastly, for the 845 news reports, we needed a more systematic way to summarize the contents. The semistructured parts like published date and journal name were gathered separately, and then we applied natural language processing algorithms to automatically extract important keywords from the news text.

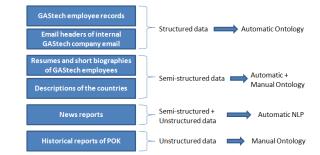


Figure 1: Data type and text analysis method.

2.1 Ontology Modeling

WO fee Pages						-	
Are their	Lost Barriel	Carl	Beet Inc.	Cont Marcel	Since many case		and Material Recorded (Resolution) and Pres.
			1				
CO-OPA, Channel		Lo Marco Destina			Provide and a second		····
+ 121.4404		- ANY OLD THE REAL					and the second
+ C2 Organisation			CRANE GARTINE	The particular of			
Cover(Name)				NetCherg 1			The intervention of the second state of the se
a California Channel			1.00	NEX. MILL			we under the charge rule; we do they of Elter inger to be user up and of going with an extension memory of 200 mills
T COOL Property of							and and an and the time takes of the group
C at second been							P. 100
Chanada taa		M-		and a local data			The Chele barrates: a total agricultural loss of alread (Call present that has 2014) for block barrathing agely etc. of therein. The p
D residentia fact			The second	Librard 1			The barrier which is appropriate areas for Tables Day for impairer. In each 1977, observe all these larges in the scansers barrier of Deriver's fault as service 1977, which, respiratory Directors, and rescriptional despirates in addition to market
+ CT indentials	PR	presenter -		- inertail			Served of thereines, that is a service, new contents, responsely, interpreted, and manifestional distances, or patients in market and Counterflowed to free action on the obtaints control mentiodities and parameter control mention of the rese, a prival parameter action of the research of the service of the obtaints control mentiodities and parameter control mention of the research of the research of the research of the service of the obtaints of the research of the research of the research of the research of the research of the resear
Dunn		property.	- ange	7 Branner 7			
O these bade has	these Processed	and the second s	- Interior	1 Branned C			In the the subscript of each double, a horizon done should pread the propider to add the become shaft.
Des		property in the	managed Flow	-	(+)		alt alterne per ortenation, loding and a painte trackper ate will estraption to tracker in the sale of a tracker in the painter is the painter the painte
C manager allow		Burden (A.	interest of the local division of the local	· source	13		
C Date for Without				and the second s			Elementar promotel ani ministra aportico.
+ 12 Parple		and a second				made 1	In the start at a second request and an other as should be and and a specific against as a mentioned for marked or which a second and a second s
- I Trans							
+ 12 Parts		and the second s				-	are and downmost offer torsee band gos fields and confirmed the presence of applications duranties will concern
COM Presented		and and				and the local division of the local division	Principle Strategies and possible Section with the Strategies (New York particular Section 2014) in constraints the section of the Strategies of the Strateg
The second second		10000				-	
+ 123 1997		10.07					and Owners anythin head with SASherh in muccus the compression and plat multi-panisheros, were give in-
- 12 1999		[edebal				market .	Big Oliver: Plana assessed Ra. mai & alter annua de caso Re. centarimator asses with Golleck (20) Rev. Selperge. 8. Secondarians in the contentry's secondaria at the Talach Resolution.
- 12 300					(mod		The state of the second second and the POC's because when is the same site of Ada - the runners a mettical with the
- 13 2001 - 13 2004		L					
+ c1 2004					Contrastitut a	and the state of the	We have a series of the second state of the se
+ C2 3004					and and have been	Sectored die	ним нам'т на должнату салтатисти нам или или тот разл вле доций и так алах. На на технат на лекона и т с пакати с запатали. Чтоткарана, на 26-колого/терник и/посеки кариса, час нас отпажние или РСК з раз
+ Colorem anti Polesi					WERE WARTS	other PCK of physics	ni is sequen ba prominante offerent pla restance of the Tairee Base Bar propinsion controlling when extense of heim is formed 2001 is retarioredness plasters and an equal. The antipolities corrected, and imaging the best
 Catriate Invalue 					proposed open		and a refer to a party and here a state from the barrier in the state of the state
+ CT Chan Servi + CT Kinds					Patrany Rd Augo	distant here	Thinks, who was suffering her reads politeria anded to for Yoster/fuer configuration, decided is they also actavate of
+ Chicagoon instant					POK. Bydaig also	1933-144.2	county diameters in an an incompanyment within one body other if a departure from the PDR, at of the other original tools contained for the totals.
· Clarge-spe solt but	THE DOCUMENT				mention of he pr	on set as per as	Column Color S and A.
- Crimi Borney					A True of Passishin	Copported to	et/Tecculmant (101) (2014)
· CZTOR THEY					Chips Ingrist's Briter		ext. Name had constraints are gradient intransic/POR Avenues. Kand incorporation in the POR matter is around its thereit
 C2 Average Sandi C2 Average Sandi 					a a para la constant		its allow an ecology publicant mentary in 2016. Yand staget runners are preside anterements rules and spart speet seeks entropy on ecology of the allow facts that the mentary allower of the second and economical added around it is
· Chinese							
· Cluster or manes					Providence in a last	a produced to	regaper and associate fore rised pride 20 a POP's reason is total a center, see your work accountability and complex
					Approval Scat pr	directed being	ta upcani il totni sud faut anchen foni politicati lagantes.
· CTMACOL THE					ABOUT MICHING	autom abobs	
+ Charaki brogenesi e	with an advectory of				Para cettera 2014		re-of passials/applied in order to see a power helt action on the conterninghan of the Trained Kines, subtle from effek POK poles
 Christeline and second and seco							rilling capitalies a 2001. Athen accelerates the 704 key abspectivel power-web complex and pointies complexing as and sendered. With other have due toget togets to an eventual a field Autophane and related other. The ender
 Cons Cur Unidentes CT Toma Materia 					Campoonth & dropp		nar an an an an ann ann ann ann ann ann
· C the farings A							
- 12 tatals intel					Tune Pinkula Fal		(Noral) De stated and dance, is come anne press press ander allafent strendt and regard we'r tre ant in shert habating is w

Figure 2: Ontology modeling interface.

Ontology is a handy method to enable semantic analysis on data. An ontology model is consisted of objects and their properties and individuals. It is useful especially when you'd like to find out indirect relations among objects or individuals. And this model is flexible to define object structure unlike database, so you can use

^{*} jhpark@astri.org

your expert knowledge more easily to address important terms for a particular data set. Most of ontology modelling was done in automatic processes. Our software reads documents, parses the structure and creates ontology classes, object properties and individuals. In addition, we provide a review and editing interface for users to manually add more information (Figure 2).

2.2 Natural Language Processing

To determine the importance of a keyword in a news report, we used Term-Frequency Inverse-Document-Frequency measure. This measure highlights unique events of a certain news contents from all the other news contents by eliminating common terms. Furthermore, as a combined word (collocation keyword) like Isia Vann can give us more meaningful and relevant information than each individual word, Isia and Vann, we used Part of Speech tags and custom dictionary to register particular people, organizations and places that we want to trace after (Figure 3). Important persons, places and organizations identified from the previous ontology analysis are saved in our custom dictionary to extract keywords from news articles more accurately.

and states and states of the	AUND [B	www.arthcine (Chevy & Anny on					
89.91		cicitus .	Table 3'rs	14.04		1 mm 201	Cent	(a) The Astro-Funt
Partiel		interest and the second	Little allen	Dependency	Date .	trol .	taryotae .	er or he sciecture
	.Mail		10011082	-Fast	2014.00	N 12	Gale-1.02	No. do a service reserves
tribe 10	the fail Name of Street or	Maple	Place	Tes	C. Bearrett	THE	(Past	
	CON MARK	Yes Ania Post Yeffine Lines				Phys. Links Provid		w WI 22 Amary 2014
		Tation like		To break of level 7 Charge of Second				In the PPOTT that is seen advices that we assume the failer advices and the provided by the set of the
	CA QU 80 70 CA QU 80 70	Assess like Lankow Lanks				Contract film for		N. TAD OPDIT. You'ld be can below new addresses. These an approximate factors indexing on another the strong for OPDITs of 27 Tax and performance of the strong for OPDITs of 27 Tax and performance of the strong for OPDITs. Strong for OPDITs of 27 Tax and performance of the strong for OPDITs of 27 Tax and performance of the strong for OPDITs. Strong for OPDITs of 27 Tax and performance of the strong for OPDITs of 27 Tax and performance of the strong for OPDITs. Strong for OPDITs of 27 Tax and performance of the strong for OPDITs of 27 Tax and performance of 27 Tax and performance of 27 Tax and performance of 28 Tax and performance o
	11112-	bine out the						Ph. Through A was conferent first over a drose instantiation was compositivities. The latter first and first first class with respectively and service and the first class with respectively and the first class with respecti
	110 110							
		Percent life		Season amounts				
				To beau of the of Y.				
		Syder-Laur				. Sinters Laters		
		Tation Issue		To beaut official 7		Totro Heren 3		
		Yby Artis Read				The Loca Paul		
		Service No.		Desting Emergen		Servery film for		
		Number-Watters Yaffrag Name				Annes Paters		
		Archist 100		To these of most 1.				
		HOMASTERS/						
								Data A Time (0.1401930) 10-47
		Arden & Ares				Drindard Kulture		Pris
								Pageo St.
								PE-Asia Falandiri D. L. RU
	01.02.0008	An proce titler		Greatery America				PT crackinelistic (319) PE crackinelistic (319)
		71010.1000		To press of mark I				[62] Addasher/1993, 279.
	010020-00	Report Die Carbon Inch		Steamy America				Re Instance and and COLUMN 1.3 MD
	01003038	Californi Bolle						Research 100
	000012021	The Approximate				The Able Full		
		Numer Sales				States & Passar		
	0000 3556	Number Tite		Distance Frences	Televis Manufel al	Directed Nat In		
					Mirashi of Talbum	Collina Paran 7		
								Brit) saysmathW. 2740
								PTZ dde mard/WH1,243
								W12 (Hall g)(802,2.80)
								P1x (64000/1,3.31)
			IR PAIRA AMPORT					Proceedings () ()
80/216		Relation little	HER.A. HOLF CO.					BLD AnderStell 2006
		794 1140	1675.A. X00008					Per 2 geneticity 1.001
1012314		The Oreand Pr	AB PURA RESOLUTION	DRAMES OF THE				Pill petmatidat(1.11)
2012014		The Galle	AREA PRACE	Tanyorye de Childa Tantor pe el Unidas		The last has		Proceeding and the second seco
2012514		The York College T		Cally and Complete		The York, Same		Rev and an and a second s
2402114		Contraction of	ANA CARA			Distance College Cor		RC strangerights (16)
2612714		The Vote	ACC. A REPORT					But respondent 1.14

Figure 3: Collocation keywords analysis.

3 VISUAL ANALYTICS TOOL

We developed a java-based visual analytic tool for the comprehensive data analysis and investigation of all available data. The main design concept was to build a consolidated view and interactive interface for users to conveniently query the basic analysis results to compose the final conclusion.

3.1 Ontology Analysis Visualization

The ontology model of country fact sheets, GAStech employee data and POK history reports are represented in a tree structure (Figure 4).

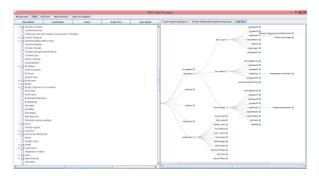


Figure 4: Ontology model tree.

3.2 Keywords Visualization

Keywords are aggregated according to the date of when the original news was published and calculated the co-occurrence count among keywords to find the linked set of keywords. Users can review the list of news in a certain day in the left window and quickly check the important events of the day in the right window (Figure 5).

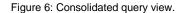


Figure 5: News keywords bubble chart.

3.3 Query and Analysis

Finally, the Query and Analysis module combines all the previous investigation results and provides users an integrated interface to dynamically search relations among organizations from their internal links and to be able to look at all the related information like published news articles for them (Figure 6). To assist the data investigation, we added a couple of more helpers such as a pop-up dialog of original news texts with highlighted keywords when users click a keyword on the screen, and more visualization options using D3 charts [4].

Background POR GAST			UKIT Der	a Processor	
Barge Onking Madels	Tock News articles Garry & Ander Self Nerged Cellsogy Medici	New New Original Optimizer Riskel	Index Collection Researchs	1	
-247	Fast Garry	Extend Annuar	Tree View	Answers: 6 (Current Quer	ry Extension Level: 0)
	9 Falaclas d Notes-POR	• MB • MB			Kespela provide Tanta different i terratio



4 CONCLUSION

Visual analytics helps us find the right information from huge and unorganized data quickly and easily. In this mini-challenge, we exerted to provide users an integrated tool to link all the available heterogeneous data in order to give them a comprehensive view of the problem and help them reach to the right answer after all.

REFERENCES

- J. Thomas and K. Cook. A Visual Analytics Agenda. *IEEE Computer Graphics and Applications*, 26:10–13, January/February 2006.
- [2] D. Keim et al. Visual Analytics: How Much Visualization and How Much Analytics? ACM SIGKDD Explorations Newsletter, 11(2):5-8, December 2009.
- [3] F. Hogenboom et al. An overview of event extraction from text. In M. van Erp et al., editors, *Proceedings of Detection, Representation,* and Exploitation of Events in the Semantic Web (DeRiVE 2011), Workshop in conjunction with the 10th International Semantic Web Conference 2011 (ISWC 2011) (Bonn, Germany, October 23, 2011), pages 48-57, 2011.
- [4] Data Driven Documents http://d3js.org/