A Network Analysis of Hermeneutic Documents Based on Bible Citations

Hajime Murai (H_MURAI@valdes.titech.ac.jp)
Department of Value and Decision Science, Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo 152-8552 Japan

Akifumi Tokosumi (AKT@ valdes.titech.ac.jp)
Department of Value and Decision Science, Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo 152-8552 Japan

Abstract

In order to handle systematic thought, we propose a new method of representing the main elements of a conceptualization in the form of a network based on the relations between the citations for frequently-used propositions within a text corpus. This method makes it possible to automatically extract the central components of a systematic thought and to analyze the relationships between these by using the clustering method. In the present study, we have constructed three networks of Christian dogma based on the writings of St. Augustine, St. Thomas Aquinas and Pope John Paul II, and have analyzed the clusters to objectively extract common elements and individual characteristics. Our network representation and analysis method help to lay the foundations for scientific research in abstract fields of human thought, such as theology. Reflecting the ways in which certain individuals perceived a particular canonical text, such as the Bible, these characterizations of key conceptualizations provide important insights into the structure of the canonical text itself. And. this method can be applied to other field where there are frequent repetitions of propositions or sentences in order to objectively analyze their meanings or interpretations.

Keywords: knowledge representation; citation analysis; Bible

Background and Goals

Background

As an aspect of higher cognitive functions, systematic thought has primarily been investigated using literature-based approaches, with texts that are usually more abstract and subjective in nature than scientific papers. However, as systematic ideas and thought influence all areas of human activity and thinking, the application of NLP (Natural Language Processing) techniques may provide a more objective understanding of systematic thought. Recently, new methods are being developed, such as the extraction of metric characteristics for text corpora (Liu, 1993; White & Mccain, 1989) specially using co-citation analysis methods (Finn, 2004; Meyer, 2004), text summarization through the automatic classification of citation information (Nanba & Okumura, 2000). Within the field of cognitive structure as well, new approaches are being developed with which to symbolize and represent concepts in the form of networks such as a Conceptual Map (Rye, 2002; Nelson & McKinney 1993). Methods that are capable of automatically extracting the main elements of a conceptualization are particularly important when handling extremely large text corpus. By utilizing these new scientific methods, we can (a) ensure the objectivity and replication of results, (b) handle large-scale data in a uniform manner, and (c) reduce information processing costs.

We believe that it is possible to analyze the abstract thoughts and value systems embodied within a text corpus with such methods. In this paper, we analyze a Christian text corpus. Traditional religions have exerted great influences on humanity throughout history. Most religions have at their core some canonical texts, with the hermeneutics, or interpretations, of the canon is also usually in text format. Thus, it is possible to represent key conceptualizations with the canonical texts through their objective analysis.

Goals

Specifically, the goals of this study are to automatically extract the main elements of a number of key conceptualizations from a large-scale religious text corpus and analyze the cluster construction of them using an objective and replicable methodology. This, in turn, will provide an objective basis for the examination of systematic thought.

Here we focus on the writings of St. Augustine and St. Thomas Aquinas, two influential Church Fathers, as well as those of Pope John Paul II to extract essential teachings of the Catholic dogma through historical transition and to identify individual characteristics of hermeneutics. Based on the patterns of the Bible citations within their writings, we created networks for frequently cited sections of the Bible, and extracted the main elements and clusters of these, in order to compare a number of key conceptualizations.

The Canon and its hermeneutic literature

The pillars of the Catholic value system are the Sacred Scripture (the Bible), the Sacred Tradition, and the Teaching Office (Paul VI, 1965; John Paul II, 1992). Of these, Bible study is the soul of sacred theology. On the other hand, the Sacred Tradition and the Teaching Office have important roles in interpreting the Bible.

Thus, Catholic dogma consists of the Bible and its interpretative literature. By the complete and systematic analysis of the interpretations of each section of the Bible and of the relationships between these, it is possible to

visualize the Catholic value system. The text data consists of writings by Pope John Paul II, posted on the Vatican Web Site (English version of Encyclicals, Motu proprio, Apostlic Exohtation, Homily, Apostlic Letter, and Speech etc. 3125 files), the writings of St. Thomas Aquinas (29 titles, including "Summa Theologica" and "Summa Contra Gentiles") and the writings of St. Augustine (29 titles, including "The city of God" and "Confessions", and 221 homilies).

Methods

First, the patterns of the Bible citations were compared for the complete corpus. Then, the relationships between the cited sections of the Bible are extracted by co-citation analysis, from which the citation networks were created. After that, the clusters of those networks are identified, and characteristics of these networks are compared.

Characteristics of the Bible citations

Comparisons of the three authors looked at citation distributions, and at which sections and books of the Bible were citations most frequently taken from.

Construction of the citation networks based on co-citation analysis

The Bible is separated into units of book, chapter and section, and numbers are assigned to these units. In this study, a citation unit is defined as the smallest section. Although there is some variation in the lengths of these sections, frequent-cited sections are generally about 1 to 2 sentences. Co-citation analysis was used to analyze these cited sections.

As Small claims, given that co-cited texts have certain similarities, the degree of similarity between texts will increase in proportion to the number of co-citations (White & McCain, 1989).

Within the Catholic hermeneutic literature, there are many Bible citations, making it possible to measure the similarity between two Bible sections in terms of the frequency of their co-citation. By connecting sections that have high frequencies of co-citation to form a citation network, it is possible to represent the complete doctrine. Although the citing unit normally used in co-citation analysis is the whole document, because these texts are divided into smaller semantic units according to Catholic tradition, this unit was used as the present analysis.

In line with one of the goals of this study—to capture the main elements of a conceptualization— we have focused on the central parts of the networks, where the elements are both frequently cited sections and for which co-citation frequency is above a certain threshold (Murai & Tokosumi, 2004).

1: For the writings of each author, the total number of citations (v_i) for a given Bible section, I, was calculated. Based on this, it was possible to select V' sections, for which v_i exceeded the citation threshold (Th_1). The each couple of the V' sections were connected if the co-citation value (e_{ij}) exceeded a co-citation threshold (Th_2).

$$\langle v_i \in V' | v_i \geq Th_1 \rangle \quad (1)$$

$$\langle e_{ij} \in E' | v_i, v_j \in V', e_{ij} \geq Th_2 \rangle \quad (2)$$

2: Connections were also made to sections that intervened between V' sections and also had co-citation values over the co-citation threshold (Th_2).

$$\left\langle v_j \in V'', e_{jk}, e_{jl} \in E'' \middle| \begin{matrix} v_j \notin V', \exists v_k, v_l \in V', \\ e_{jk} \geq Th_2, e_{jl} \geq Th_2, \end{matrix} \right\rangle \quad (3)$$

3: Creating the network by connecting V' E' V'' E''.

Extraction of the clusters within the citation networks

Networks are usually composed of some clusters. We can understand the characteristics of networks by extracting the clusters and analyzing these mutual relationships.

There are many clustering algorisms. In this paper we extract clusters by shifting the co-citation threshold (Th_2). The first merit is that the clusters reflect the strength of co-citations. Normal methods that extract some type of cliques reflect only the network topology. The second merit is that we can change the size of the clusters by shifting the co-citation threshold, making it easier to extract clusters of appropriate sizes.

Results

Citation and Co-citation Patterns

Table 1 presents all citations, all cited sections, all co-citations, and all co-cited section pairs, together with all sections that include the Bible citations and the average citations per unit for the three authors. While the writings of St. Thomas include more various citations, those of John Paul II have little diversity in terms of the total citations.

Table 1: Total citations.

	Augustine	Thomas	John Paul II
All citations	22674	36015	32166
All cited sections	8645	11821	8851
All co-citations	215824	800457	643708
All co-cited sections	189353	754201	508118
All sections	3268	2393	3444
Average citations	6.938	15.050	9.340

Table 2 shows the most frequently cited books of the Bible. The numbers of each book are the ratios of the

citations of that book divided by total sections of that book. Of these, 14 books are common to three authors (Bold). This result indicates a similarity in their patterns of citation. The Gospel according to John is the most cited Gospel. And, John Paul II cites more sections from the Gospels than the others. However, analysis of the most frequently cited sections, (Table 3), shows that common sections are rather rare. There are only 5 common sections (Bold: Jn1:14, Jn14:6, Ph2:7, Ph2:8, Rm5:5) among the top 40.

Citation Networks and Clusters

In comparing the three authors, the number of V' elements and the size of the maximum connected component were adjusted by the thresholds. For St. Augustine, Th_1=25, Th_2=4 and for St. Thomas Aquinas, Th_1=27, Th_2=5 and for Pope John Paul II, Th_1=41, Th_2=5. Figures 1, 2, and 3 show the respective maximum connected partial graphs. The numbers within the [ ] are cited number, and the values V' E' V'' E'' are indicated as follows:

The clustering threshold is selected so that the numbers of clusters in each network is similar (in this paper, clusters are 4 or 5), and thresholds were for St. Augustine 6, for St. Thomas 7, and for Pope John Paul II, 7. The results of clustering are represented in Figures 4, 5, and 6.

Table 2: Top 20 most frequently cited books.

Augustine		Thomas		John Paul II	
Rm	4.958	**Rm**	5.155	**Ep**	6.568
Ga	3.732	**Ep**	4.761	**Jn**	5.058
Ph	3.615	**1Co**	4.746	**Ph**	3.885
1Co	3.336	**Ph**	4.192	Lk	3.514
Ep	3.097	**Ga**	3.718	**Mt**	3.415
Mt	2.498	**1Jn**	3.667	**Ga**	3.396
1Jn	2.371	Jm	3.389	**Col**	3.368
Col	2.316	**1Tm**	3.336	**1Jn**	3.305
1Tm	2.283	**1P**	3.257	**1Co**	3.096
Jn	2.222	**Jn**	3.107	**Rm**	2.988
2Co	2.082	Heb	3.066	**1P**	2.505
2Tm	1.542	**Col**	3.053	Heb	1.749
Tt	1.478	**2Co**	2.941	**2Tm**	1.735
Ml	1.473	**Mt**	2.899	Ps	1.621
Jm	1.324	**2Tm**	2.313	Mr	1.600
Gn	1.223	**Tt**	2.304	Ac	1.595
Ac	1.212	Ml	2.091	Rv	1.489
1P	1.210	2P	1.934	**1Tm**	1.416
2Th	1.149	Ws	1.628	**2Co**	1.367
2P	1.148	1Th	1.517	**Tt**	1.196

Table 3: Top 40 citations for all sections of each books.

Augustine		Thomas		John Paul II	
Mt6:12	84	**Jn1:14**	66	Mt28:20	123
Jn1:14	70	**Rm5:5**	55	**Jn14:6**	112
Rm5:5	68	Rm8:15	50	Jn17:21	111
Jn1:1	66	Ep3:17	48	Jn3:16	103
Rm5:12	64	Ga4:4	48	Mt28:19	92
1Co13:12	56	Heb11:6	47	Heb13:8	83
Jn1:3	55	Jn1:17	43	Lk4:18	82
Rm7:25	52	Rm1:20	43	Jn10:10	76
Ga5:6	51	Heb1:3	42	Mt25:40	74
1Tm2:5	47	**Ph2:7**	41	**Jn1:14**	69
Ga5:17	44	**Ph2:8**	41	Gn1:26	68
Ph2:6	43	Rm8:29	40	**Rm5:5**	66
Rm7:23	43	1Co13:12	39	Ga4:4	64
Rm7:24	40	Ep5:2	39	**Ph2:8**	63
Ph2:7	39	Jn17:3	39	Mt5:14	63
1Jn1:8	38	1Co15:10	37	Gn1:27	59
Ws9:15	38	1Jn4:16	37	Jn13:1	59
Jn14:6	36	Jn1:18	37	Mt5:13	59
Mt5:8	34	1Jn3:2	35	Lk5:4	58
Mt6:13	34	Is53:7	34	**Ph2:7**	56
Rm7:18	34	1Co4:1	33	Jn20:22	54
Jn1:9	32	**Jn14:6**	33	Ac1:8	54
1Co1:31	31	Ph3:21	33	Gn1:28	51
1Co4:7	31	Gn1:26	32	Lk1:38	50
1Jn3:2	31	Rm6:23	32	Ep2:14	49
Ac2:4	31	Rm8:3	32	Mt25:35	49
Rm1:17	30	1Tm2:4	31	Jn16:13	46
Mt25:41	29	Jn3:5	31	Jn20:21	46
Jn3:5	28	Rm5:12	31	Mr1:15	46
Rm10:3	28	Mt1:21	30	Mt25:36	45
Jn10:30	27	Rm10:10	30	1p3:15	45
Jn15:5	27	Ep2:8	29	Jn2:5	44
Rm1:20	27	Ex20:12	29	Ac1:14	43
Ph2:8	26	Mt5:16	29	Ac2:42	42
Rm7:22	26	Mt11:29	28	Jn10:11	42
1Co3:7	25	Ws8:16	28	Jn19:27	41
2Co4:16	25	Heb11:1	27	Ga4:6	41
Ac2:3	25	Mt19:21	27	Jn15:5	41
Gn1:27	25	Rm6:4	27	Lk22:32	41
				Lk4:19	41

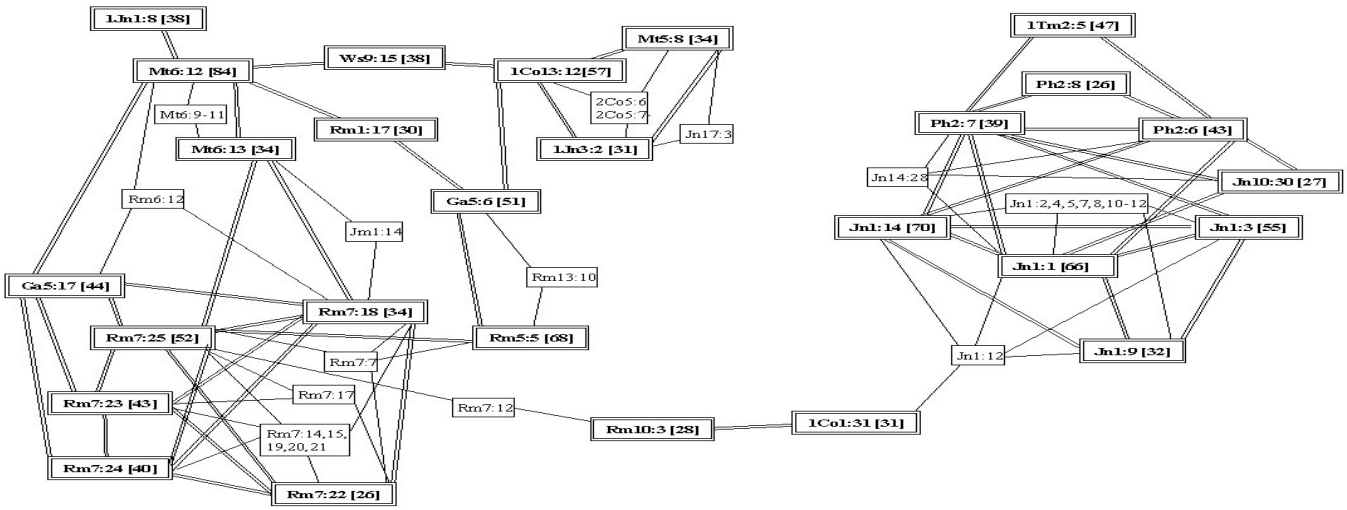

Figure 1: Citation Network for St. Augustine.

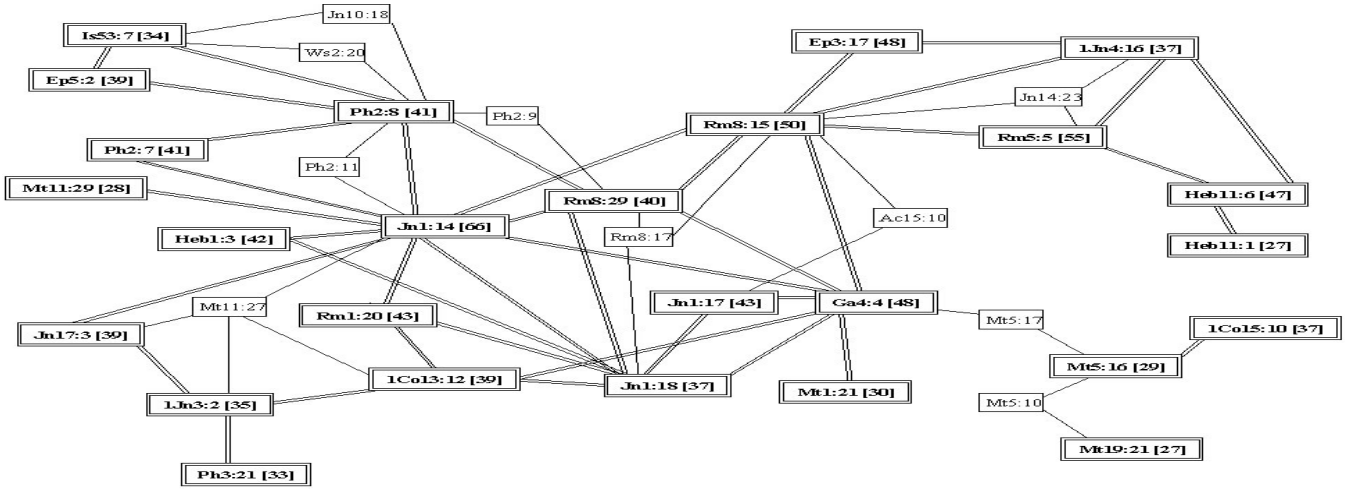

Figure 2: Citation Network for St. Thomas Aquinas.

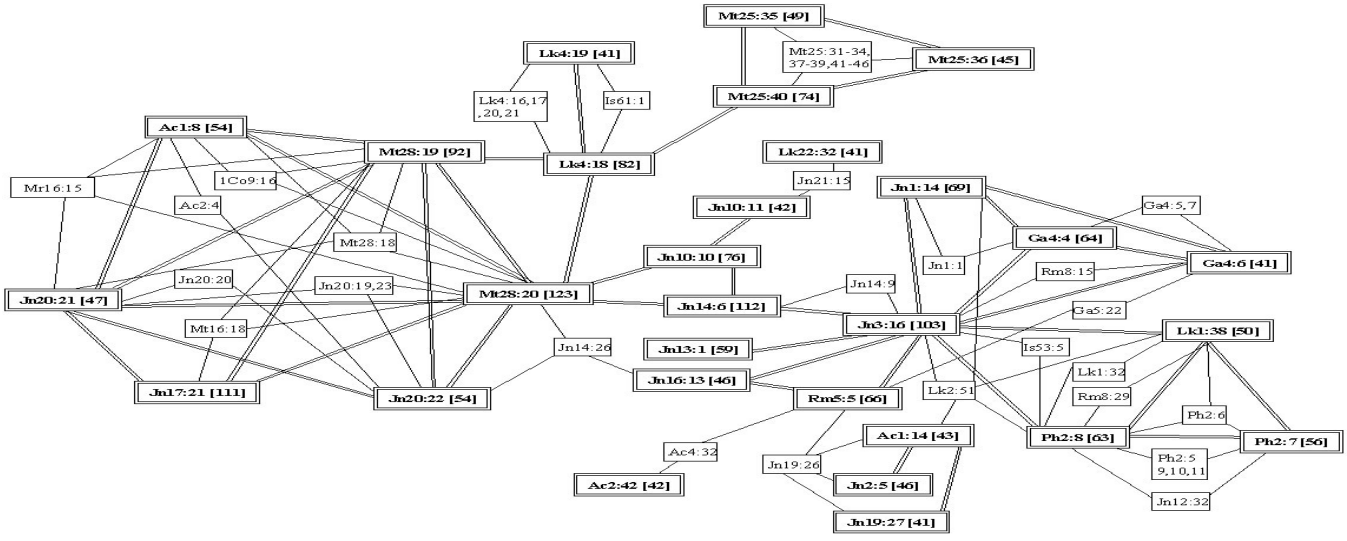

Figure 3: Citation Network for Pope John Paul II.

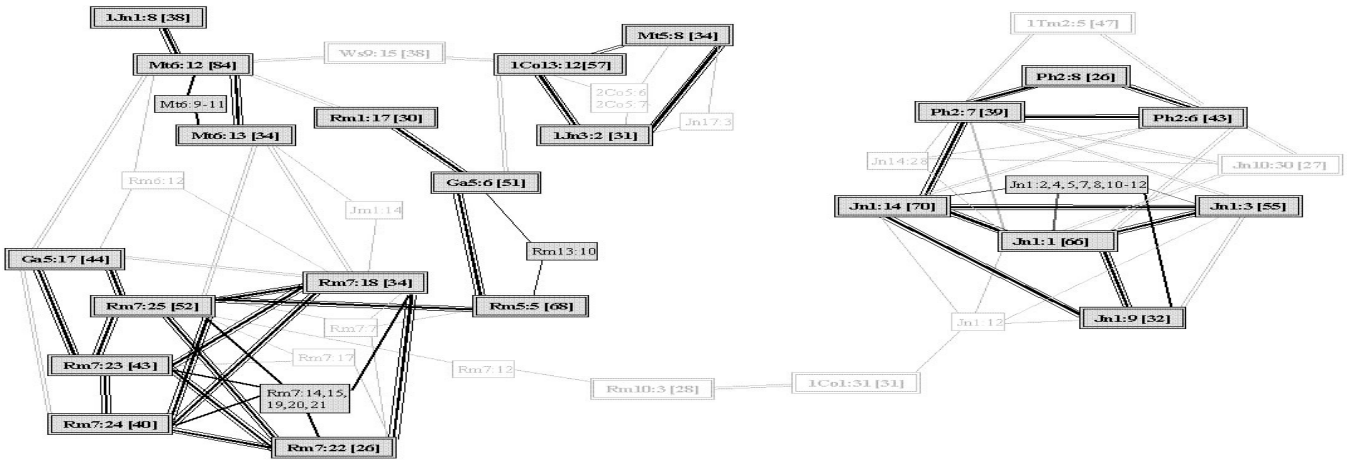

Figure 4: The Citation Network for St. Augustine after Clustering Analysis.

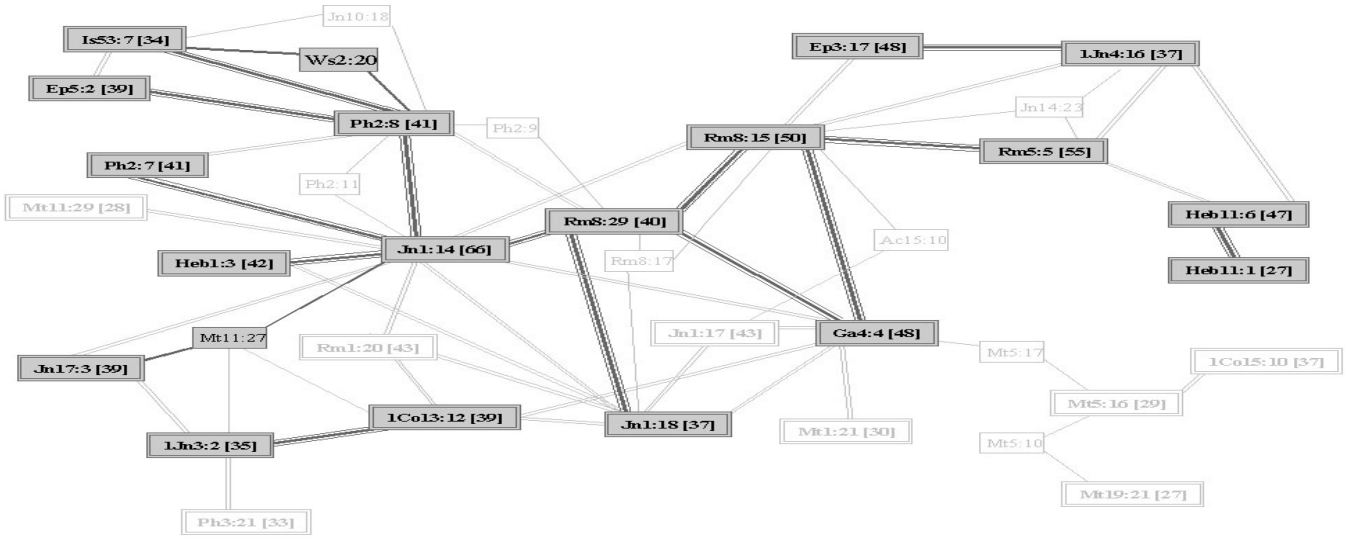

Figure 5: Clustering Result of Citation Network for St. Thomas Aquinas.

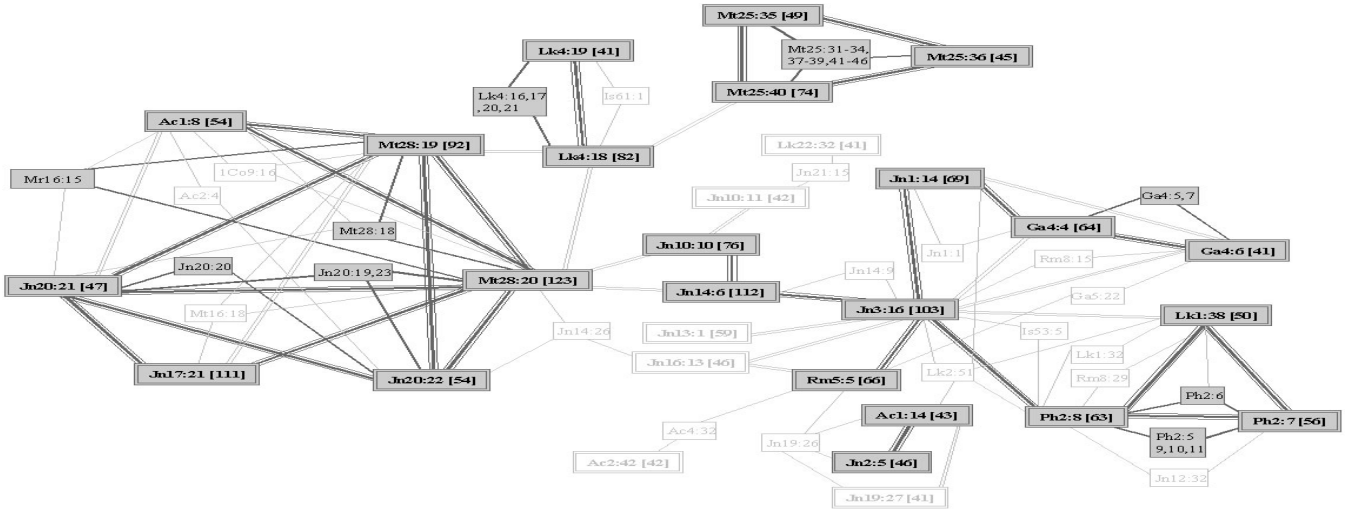

Figure 6: Clustering Result of Citation Network for Pope John Paul II.

Discussions

Individual Characteristics

St. Augustine's network consists of two relatively separate parts; one including Matthew 6:12-13 and Rome 7:18-25 and one covering John 1:1-14 and Philippians 2:6-8. Although his two main themes are incarnation and the struggle with sin, they are only weakly related.

There are stronger connections between the clusters within St. Thomas's network, where the central concepts are concerned with incarnation (John 1:14, Philippians 2:7-8, Ephesians 5:2), the Trinity (Rome 5:5, 8:15, 29, Galatians 4:4), and love and faith (Ephesians 3:17, 1 John 4:16).

The network for Pope John Paul II has a large cluster concerned with spreading the Gospel (Matthew 28:19-20, John 17:21, 20:20-21, Acts 1:8). This may reflect the presence within the corpus of highly authoritative texts that are probably more representative of Vatican opinion.

Common Characteristics

Common elements in the teachings of the Catholic Church are a constant emphasis on Jn1:14, Jn14:6, Ph2:7, Ph2:8, Rm5:5. According to the clustering results, Jn14:6, Ph2:7 and Ph2:8 compose common clusters in three author's networks. These sections are about Incarnation.

Jn14:6 is "I am the way and the truth and the life. No one comes to the Father except through me." And Rm5:5 is "hope does not disappoint, because the love of God has been poured out into our hearts through the holy Spirit that has been given to us." These sections are included in different clusters on each network. Pope John Paul II includes both within a cluster about Incarnation, but St. Augustine includes Rm5:5 within a cluster that focuses on sin. So, we may regard these two interpretations as being different.

From the citation analysis and citation networks, we may conclude that the most important teaching of the Catholic Church common throughout its history is the Incarnation of Christ.

Conclusions

By representing the frequently-used elements of the network structure, it is possible to objectively perceive the complete systematic framework of this complex theology. This is functionally similar to automatic summarization for meaning.

Moreover, the network structure is also useful for extracting and numerically analyzing semantic differences in fields where complex interpretations are required.

This paper demonstrates that our new method is effective by comparing abstract thoughts and extracting common and unique elements in those abstract thoughts objectively. Our network representation and analysis provides a scientific basis for research in more abstract fields, such as theology.

To the extent that the extracted structures are reflecting the semantic structures perceived by the interpreter, it also provides a means of capturing the inner structure of the Bible. Accordingly, this network construction and clustering method can be applied to other field where there are frequent repetitions of propositions or sentences.

References

Douglas L. Nelson & Vanesa M. McKinney, (1993) Implicit Memory Effect of Network Size And Interconnectivity On Cued Recall, *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 747-764.

Finn R. Forsund & Nikias Sarafoglou, (2004) The tale of two research communities: The diffusion of research on productive efficiency, International Journal of Production Economics, In Press, Corrected Proof, Available online.

Hajime Murai & Akifumi Tokosumi, (2004) A Network Representation of Hermeneutics Based on Co-citation Analysis, *WSEAS Transactions on Information Science and Applications*, l1, 6, 1513-1517.

Hidetsugu Nanba & Noriko Kando & Manabu Okumura, (2000) Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation, *The American Society for Information Science (ASIS) the 11th SIG Classification Research Workshop, Classification for User Support and Learning,* 117-134.

James Rye & Peter Rubba, (2002) Scoring Concepts Maps: An Expert Map-Based Scheme Weighted for Relationships, *School Science and Mathematics*, 33-44.

Liu, M, (1993) Progress In Documentation The Complexities of Citation Practice: A Review of Citation Studies, *Journal of Documentation*, 49, 4, 370-409.

Martin Meyer & Tiago Santos Pereira, (2004) Olle Persson and Ove Granstrand, The scientometric world of Keith Pavitt: A tribute to his contributions to research policy and patent analysis, *Research Policy*, 33, 9, 1405-1417.

Pope John Paul II, (1992) Catechism of the Catholic Church, http://www.vatican.va/archive/ENG0015/_INDEX.HTM.

Pope Paul VI, (1965) Documents of the II Vatican Council Dei Verbum, http://www.vatican.va/archive/hist_councils/ii_vatican_council/documents/vat-ii_const_19651118_dei-verbum_en.html.

White, H. D. & Mccain, K. W., (1989) Bibliometrics, *Annual Review of Information Science and Technology*, 24, 119-186.