PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Giancarlo Sperlì

XXX Cycle

Training and Research Activities Report – First Year

Tutor: Antonio Picariello

Student

I graduated in Computer Engineering; currently, I am attending the first year of PhD in Information Technology and Electrical Engineering - ITEE- XXIX Cycle at the University of Naples Federico II, under the supervision of Prof. Antonio Picariello. I was awarded a MIUR research grant.

Study Activities

Courses

Module	Type	Professor	Date	H	CFU
The Entrepreneurial Analysis of Engineering Research projects	Ad hoc	Iandoli	20/02/2015	15	3
PROJECT MANAGEMENT PER LA RICERCA	Ad hoc	Capaldo	13/03/2015	16	3
Modelli, metodi e software per l'ottimizzazione	Ad hoc	Sforza	27/05/2015	18	4
Three core issues for the internet: things, security and economics	Ad hoc	Romano	20/02/2015	8	2
Designing and writing scientific manuscripts for publication in english language scholarly journals, and related topics	Ad hoc	Parker	17/06/2015	12	3

Seminars

Module	Type	Professor	Date	H	CFU
Social Signal Processing:,understanding social interactions,through nonverbal behavior analysis	Ext	Vinciarelli	05/05/2015	2	0,4
Agents with Truly perfect Recall	Ext	Bulling	28/04/2014	1	0,2
Colloquium on Robotics	Ext	Siciliano	21/04/2015	5	1
Partial Possibilistic Regression Path Modeling	Ext	Romano	20/04/2015	2	0,4
Affidabilità di dispositivi e moduli elettronici di potenza	Ext	Castellazzi	24-26/03/2015	6	1,2
A new look at Electro-magnetic Induction	Ext	Romano	19/03/2015	2,5	0,5
The iCub project: An open platform for research in robotics & Artificial Intelligence	Ext	Metta	18/03/2015	1,5	0,3
Answering queries over inconsistent databases	Ext	Murano	18/03/2015	1	0,2
Site Reliability Engineering at Google	Ext	Manzillo	27/11/2014	3	0,6
Verifica e Validazione di sistemi Safety Critical	Ext	Tramontana	16/12/2014	2	0,4
Efficient service distribution in next generation cloud	Ext	Tulino	10/02/2015	4	0,8
Linked Open Data-enabled Strategies for Top-N Reccomendations	Ext	Basile	05/02/2015	1,5	0,3
State of the art in Power Converters for high voltage DC trasmission systems	Ext	Ladoux	28/01/2015	2	0,4
Summer School "Sodata2015"	Ext	Picariello	12/06/2015	24	4,8

	Credits year 1							
	Estimated	1 bimonth	2 bimonth	3 bimonth	4 bimonth	5 bimonth	6 bimonth	Summary
Modules	20	0	5	3	7	0	0	**15(*)**
Seminars	10	1	1,5	3,8	5,2	0	0	**11,5**
Research	30	9	3,5	3,2	3	10	10	**38,7**
	60	10,0	10,0	10,0	15,2	10,0	10,0	**60**

(*) Waiting for perform final exam of "Semantic web reasoners: struttura, uso e ottimizzazioni" (5 CFU) held by Prof. Bonatti

Research Activity

According to a report from International Data Corporation (IDC)[1], the amount of data, created and copied, grew nine times in the last five years, reaching the size of 1,8 ZB (1021 B). This large amount of data, called Big Data, can include different kind of information such as transactional data, warehoused data, metadata, and other data residing in large files related to Media/entertainment, healthcare, and video surveillance environments. Moreover, another kind of data sources are social media solutions such as Facebook, Foursquare, and Twitter.

In [2], McKinsey has analyzed the values obtained by using of Big Data in medical environment: if big data could be relatively and effectively utilized to improve efficiency and quality, the potential value of the U.S medical industry gained through data may surpass USD 300 billion, thus reducing the expenditure for the U.S.healthcare by over 8 %; retailers that fully utilize big data also be utilized to improve the efficiency of government operations, such that the developed economies in Europe could save over EUR 100 billion (which excludes the effect of reduced frauds, errors, and tax difference)[3].

According to a HACE Theorem, Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control and seeks to explore complex and evolving relationships among data.

These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. Thus, it grew in importance the need to store and retrieve large amount of data. For this reason, multi-dimensional and high-dimensional indexing in decentralized peer-to-peer (P2P) networks have received extensive research attention. Naturally, most such methods are tree-based and the data space is hierarchically divided into smaller subspaces (regions), such that the higher level data subspace contains the lower level subspaces and acts a guide in searching. These methods can be data-partitioning based, where data subspaces are allowed to overlap (eg. R-tree) or space-partitioning based, where data subspaces are disjoint (eg. kd-tree)

These method can be classified into three categories:

1. Tree-based: There have been a number of indexing data structures suggested to handle high-dimensional data: R-tree, Kdtree, X-tree, SS-tree, M-tree, Quadtree, etc. These methods exhibit logarithmic search cost, but face a serious limitation. Peers that correspond to nodes high in the tree can quickly become overloaded as query processing must pass through them. In centralized indices this was a desirable property because maintaining these nodes in main memory allow the minimization of the number of I/O operations. In distributed indices it is a limiting factor leading to bottlenecks. Moreover, this causes an imbalance in fault tolerance: if a peer high in the tree fails than the system requires a significant amount of effort to recover. Some proposal are been made: MIDAS[4], composing by physical and virtual nodes used for load balancing and fault tolerance purpose, MD-HBase[5], scalable data management system for location based sevices and quadtree over a range partitioned Key-value store index in peer-to-peer networks[6], where peer have responsibilities for region of spaces.
2. DHTS-based: These approaches are based on distributed hash tables (DHTs) and they employ a globally consistent protocol to ensure that any peer can efficiently route a search to the peer that has the desired content, regardless of how rare it is or where it is located. A DHT system provides a lookup service similar to a hash table; (key, value) pairs are stored in a DHT, and any participating node can efficiently retrieve the value associated with a given key. Responsibility for maintaining the mapping from keys to values is distributed among the nodes, in such a way that a change in the set of participants causes a minimal amount of disruption. This allows a DHT to scale to extremely large numbers of nodes and to handle continual node arrivals, departures, and failures.
3. Skiplist-based: Skip Graphs [7] and SkipNet [8] are two skip-list based structured P2P systems. Skip Graphs and SkipNet maintain O(logN) neighbors in their routing table. For each node, the neighbor at level h has the distance of 2h to this node. Finally, SCRAP [9], ZNet [10], employ a space filling curve , such as Hilbert or z-curve, to map the multidimensional space to a single dimension and then use a conventional system to index the resulting space.

One of the main sources of Big Data are today On-line Social Networks (OSNs). A OSN can be seen as a big data research field given its intrinsic characteristics, described by [11], in fact:

- **Data Availability**: the large amount of data produced from users belonging to OSNs and change rate of data are the key characteristics of Big Data
- **Multiple Authorship**: in an OSN different authors can publish different kind of multimedia objects, making OSN a large repository of multimedia data.
- **Agent interaction**: in an OSN it's possible identify different kind of relationship, given that each user can be interact with each other.
- **Temporal Dynamics**: a key role is playing by a temporal dimension of relationship established in an OSN.
- **Instantaneity**: users belonging to OSNs produce multimedia data in response to internal and external stimuli.
- **Ubiquity**: Exploiting continuous increasing of technology, users can be publish multimedia data anywhere and any time.

This information contains both social and multimedia contents that will be exploited in different applications, such as viral marketing, social recommendation, influence analysis and so on.

Different approaches have been developed in order to model this information based on the type of application identified. For influence analysis, Kempe et al.[12] propose an approach based on graph, whose node are the users and edges represent the influence exerted from user to another user belonging to same networks. In [13], the authors exploits this kind of network to identify the experts between users. Taganelli et al.[14] develop a similar approach to identify silent members, called lurkers, into the network. A different approach are been developed by Qui et al [15], that propose a tri-partite graph, whose vertices are users, tag and multimedia objects whereas edge set are composed by directed link from user to its interested multimedia object and by undirected link that interconnect tag and multimedia object, in order to cluster object, exploiting social information . An approach that exploits hypergraph theory has been developed by Liu et al [16]. The authors define a graph, in which nodes are heterogeneous vertices and only the arc interconnecting user, tag and image vertex are model through an hyperarc to recommendation field. Eventually, Bu et al [17] propose an approach based on hypergraph network in order to develop a music recommendation exploiting both social and acoustic based information.

References

[1] https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
[2] Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute
[3] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: A survey." Mobile Networks and Applications 19.2 (2014): 171-209.
[4] S. Naz, M. Naeem, and A. Qayyum, “Performance evaluation of index schemes for semantic cache,” International Journal of Information Technology and Computer Science (IJITCS), vol. 5, no. 4, p. 40, 2013.
[5] S. Nishimura, S. Das, D. Agrawal, and A. E. Abbadi, “Md-hbase: a scalable multi-dimensional data infrastructure for location aware services,” in Mobile Data Management (MDM), 2011 12th IEEE International Conference on , vol. 1. IEEE, 2011, pp. 7–16.
[6] E. Tanin, A. Harwood, and H. Samet, “Using a distributed quadtree index in peer-to-peer networks,”The VLDB Journal, vol. 16, no. 2, pp. 165–178, 2007.
[7] J. Aspnes and G. Shah, “Skip graphs,” ACM Transactions on Algorithms (TALG) , vol. 3, no. 4, p. 37, 2007.
[8] N. J. Harvey, M. B. Jones, S. Saroiu, M. Theimer, and A. Wolman, “SkipNet: A Scalable Overlay Network with Practical Locality Properties.” In USENIX Symposium on Internet Technologies and Systems ,vol. 274. Seattle, WA, USA, 2003
[9] B. Y. Zhao, L. Huang, J. Stribling, S. C. Rhea, A. D. Joseph, and J. D. Kubiatowicz, “Tapestry: A resilient global-scale overlay for service deployment,” Selected Areas in Communications, IEEE Journal on , vol. 22, no. 1, pp. 41–53, 2004.
[10] A. Rowstron and P. Druschel, “Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems,” in Middleware 2001 . Springer, 2001, pp. 329–350
[11] D. B. Kurka, A. Godoy, and F. J. Von Zuben, “Online social network analysis: A survey of research applications in computer science,” arXiv preprint arXiv:1504.05655, 2015.
[12] Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.
[13] Zhang, Jing, Jie Tang, and Juanzi Li. "Expert finding in a social network." Advances in Databases: Concepts, Systems and Applications. Springer Berlin Heidelberg, 2007. 1066-1069.

[14] Tagarelli, Andrea, and Roberto Interdonato. ""Who's out there?" Identifying and ranking lurkers in social networks." Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013.
[15] Qi, Guo-Jun, Charu C. Aggarwal, and Thomas S. Huang. "On clustering heterogeneous social media objects with outlier links." Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.
[16] D. Liu, G. Ye, C.-T. Chen, S. Yan, and S.-F. Chang, "Hybrid social media network," in Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 659–668.
[17] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He, "Music recommendation by unified hypergraph: combining social media information and music content," in Proceedings of the international conference on Multimedia. ACM, 2010, pp. 391–400.

Products

F. Amato, A. De Santo, F. Gargiulo, V. Moscato, F. Persia, A. Picariello, G. Sperli: *"A Novel Approach to Query Expansion based on Semantic Similarity Measures" (344-353), DATA 2015:*

F. Amato, A. De Santo, V. Moscato, A. Picariello, D. Serpico, G. Sperli*: "A Lexicon-Grammar Based Methodology for Ontology Population for e-Health Applications"(521-526) CISIS 2015, Blumenau, Brazil*

F. Persia, D. D'Auria, G.Sperlì, A. Tufano: *"A prototype for Anomaly Detection in Video Surveillance Context"(517-528) Somet 2015, Naples, Italy:*

A. D'Acierno, F. Gargiulo, V.Moscato, A. Penta, F. Persia, A. Picariello, C. Sansone, G. Sperlì *"A Multimedia Summarizer integrating Text and Images" (21-33) IIMSS 2015, Sorrento, Italy*

Flora Amato, Aniello De Santo, Vincenzo Moscato, Fabio Persia, Antonio Picariello, Silvestro Roberto Poccia and Giancarlo Sperlì. "*A structure-based approach for ontology partitioning*", SEBD 2015, Gaeta, Italy