Prosiding Seminar Nasional MIPA 2016
"Peran Penelitian Ilmu Dasar dalam Menunjang Pembangunan Berkelanjutan"
Jatinangor, 27-28 Oktober 2016
ISBN 978-602-72216-1-1

SparkR: Pemograman R pada Data Besar

Zulhanif*, I Gede Mindra Jaya, Bertho Tantular

Departemen Statistika, Universitas Padjadjaran, Bandung
*E-mail: dzulhanif@yahoo.com

Abstrak

Software R merupakan *software* Statistika yang pertama kali diluncurkan pada tahun 1993, dari University of Auckland, New Zealand, telah berhasil mendapatkan apresiasi dari berbagai kalangan baik profesional maupun akademisi. Peningkatan penggunaan R sendiri disebabkan karena R berbasis *open source* dan memiiki tingkat flexibilitas yang tinggi sehingga dalam waktu yang singkat memiliki komunitas pengguna yang besar. Pada makalah ini akan dijelaskan salah satu alternatif solusi untuk menangani data besar dengan menggunakan sparkR. sparkR sendiri merupakan library pada R yang merupakan *front end* dari Apache Spark, sparkR memungkinkan pengguna menggunakan *dataframe* terdistribusi layaknya database pada Apache Spark dalam pemprosesan data dalam skala besar.
Kata kunci: Data Besar, R Sotware, Apache Spark dan SparkR

1. Pendahuluan

Tren perkembangan terbaru dalam analisis data besar, menunjukkan peningkatan yang signifikan dalam hal analisis interaktif data besar (Barnett et al., 2013). Menanggapi tren ini, sejumlah kalangan akademik (Kornacker et al., 2015) maupun industri mencoba memecahkan masalah tersebut dengan mengembangkan perangkat lunak yang mampu mendukung perkembangan data besar. Analisis data besar juga tidak hanya terbatas pada database, akan tetapi berkembang juga pada data survei, seperti data Survei Sosial Ekonomi Nasional (Susenas). Perkembangan ini menunjukkan bahwa selain pemrosesan *query relasional*, Analis data juga sering melakukan analisis lebih lanjut untuk dapat menarik kesimpulan dari suatu data, salah satu *tools* yang sering dipergunakan untuk menganalisis data adalah software R. *Software* R sendiri dipilih karena berbasis *open source* dan didukung oleh pustaka atau library yang sangat lengkap dalam hal analisis data. Secara umum pengolahan data didalam R berbasis data terstruktur yang dikenal sebagai *data frame*. *Data frame* sendiri merupakan *object* didalam R yang menampung data *numeric* maupun *non numeric* secara bersamaan. Akan tetapi penggunaan R dalam analisis data, dibatasi oleh jumlah memori yang tersedia pada mesin tunggal (*single threaded*). Selanjutnya penggunaan R sebagai *single threaded* sering tidak praktis untuk penanganan dataset. yang besar. Beberapa penelitian berusaha mengatasi permasalahan keterbatasan ini melalui dukungan input output (I/O) (Zhang et al., 2010) yang lebih baik, integrasi dengan Hadoop (Das et al., 2010) dan merancang pendistribusian *R runtimes* (Venkataraman et al., 2013) yang dapat diintegrasikan dengan mesin *database management system* (DBMS) (Prasad et al., 2015). Penanganan data besar sendri dapat dilakukan degan berbagai software statistik salah satunya adalah R melalui SparkR. SparkR sendiri merupakan *frontend* R untuk Apache Spark yang secara luas telah digunakan (Sparks, 2014) pada mesin komputasi *cluster*. Adapun sejumlah manfaat dengan merancang sebuah *frontend R* yang terintegrasi dengan Spark adalah dukungan pustaka Spark berisi untuk menjalankan query SQL (Armbrust et al., 2015) dan *machine learning* (Meng et al., 2015) serta grafik analisis (Gonzalez, 2014). Dalam penelitian ini, akan dibahas bagaimana menggunakan program R melalui sprakR dalam menangani data besar pada *database* pendistribusian beras miskin (RASKIN) di Jawa Barat.

2. Metode

Metode pengelolaan data besar dengan SparkR pada dasarnya menggunakan fungsi atau *library* pada program R sehingga tidak memerlukan perubahan Komponen R. Inti dari SparkR adalah penggunaan *data frame* terdistribusi yang memungkinkan pengolahan data terstruktur dengan sintaks R (Wickham,H. dan Francois,2015). Sedangkan untuk meningkatkan kinerja dari dataset besar, SparkR melakukan operasi *query optimizer* (Armbrust et al., 2015) pada *data frame*. SparkR sendiri awalnya dikembangkan di AMPLab, UC Berkeley dan telah menjadi bagian dari proyek Spark Apache Spark sebelumnya. Sampai saat ini SparkR adalah sebuah proyek aktif dengan lebih dari 40 kontributor.

2.1 Metode Pengumpulan Data

Penggunaan SparkR dalam penelitian ini menggunakan data paten yang bersumber pada *database patent* internasional yang dapat diakses di https://worldwide.espacenet.com/, sedangkan kata kunci pencarian *database patent* ini adalah patent yang berkaitan dengan penyakit. Adapun tujuan dari analisis data paten adalah identifikasi tema-tema dari teknologi-teknologi penanganan penyakit dan produk-produk obat-obatan. Disamping itu hasil analisis lanjut "IPC mapping" akan membantu peneliti memilih area-area

ISBN 978-602-72216-1-1

teknologi penanganan penyakit untuk *innovative thinking.*

2.2 Metode Analisis Data

Dokumen paten yang tersimpan dalam data base internasional mengandung komponen informasi bibliographic yang tersusun dalam elemen-elemen informasi (*title, patent number, publication date, inventor, applicant, classification* dan *abstrac*), disebut *field*. Konsep analisis data paten dalam penelitian ini adalah analisis data pada suatu set bibliographic data paten. Salah satu metode yang paling banyak digunakan untuk menganalisis data tersebut adalah analisis bibliometrika yang sudah banyak digambarkan pada beberapa publikasi (Dou, 2004). Bibliometrika adalah instrument pengukuran dan analisis berdasarkan pemakaian teknik-teknik statistik yang bertujuan untuk membantu dalam membandingkan dan memahami suatu set besar elemen-elemen bibliographic melalui berbagai korelasi (Rostaing, 1996). Pada bagian data paten yang tidak terstruktur seperti judul dari patent, metode analisis yang dipergunakan adalah analisis *clustering*. *Clustering* adalah suatu proses mengelompokan data ke dalam sebuah kelas atau cluster, dimana objek yang berada di dalam klaster memiliki tingkat kemiripan yang tinggi satu sama lainnya tetapi memiliki tingkat ketidakmiripan yang tinggi dengan objek di klaster lain (Handan Kamber, 2006). Dalam beberapa aplikasi, clustering juga dapat disebut sebagai segmentasi data karena *clustering* mempartisi suatu set data yang besar kedalam sebuah kelompok berdasarkan kemiripannya. *Clustering* merupakan *unsupervised learning* technique. *Clustering* dapat diaplikasikan di banyak bidang seperti analisis DNA, marketing studies dan *text mining* (Singh dan Chauhan, 2011). *Clustering* sangat berguna untuk memperoleh pola dan struktur dari sebuah set data yang besar. Secara umum, metode *clustering* dibagi menjadi dua kelompok, yaitu metode *hierarki* dan metode *non-hierarki*. Pada penelitian ini akan dipergunakan metode *non-hierarki* Bisection K-means. Adapun langkah awal dalam pemrosesan data besar dengan SparkR pada OS Windows sbb:

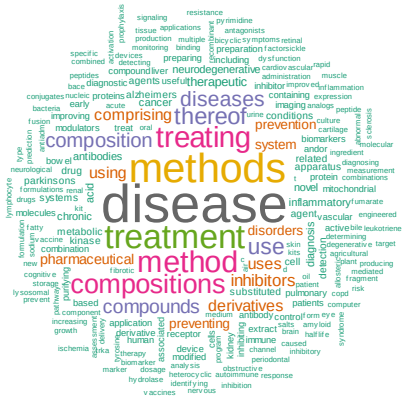
1. Melakukan Instalasi Spark dari http://spark.apache.org/
2. Lakukan *Built Package* Spark
3. Jalankan IDE SparkR pada command windows.
4. Jalankan code R Bisection K Means pada console R.

Algoritma Bisection K-means pada dasarnya merupakan pengembangan dari algoritma K-means, Penerapan algoritma ini secara luas diterpakan pada proses pengklasteran teks (Steinbach et al., 2000). Secara lengkap algoritma Bisection K-means adalah sbb:

1. Tetapkan jumlah *cluster k.*
2. *Split* data menjadi 2 *cluster* dangan metode K-means.
3. Pilih cluster yang akan di*split* berdasarkan *sum square error* (SSE) terkecil.
4. Ulangi langkah 2 sampai *k cluster* tercapai.

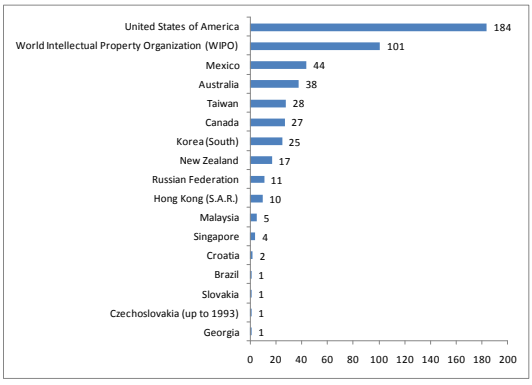
3. Hasil dan Pembahasan

Proses awal analisis dilakukan dengan melakukan tahapan pembersihan data teks yang dalam hal ini adalah judul patent melalui tahapan *cleaning.* Setelah tahapan *cleaning* data text sudah dilakukan langkah selanjutnya adalah dengan membuat matriks kemunculan kata berdasarkan dokumen patent, proses ini dilakukan dengan bantuan software R. Hasil analisis awal menunjukkan *term* data memiliki tingkat kejadian yang paling tinggi jika dibandingan dengan *term* kata lainnya hal ini dapat dilihat pada Gambar 1.



Gambar 1 *Wordcloud Term*

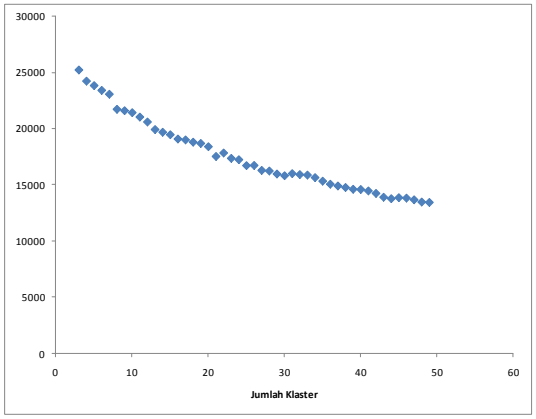
Selanjutnya jika ditinjau dari negara jumlah aplikasi patent dapat dilihat pada Gambar 2.



Gambar 2 Data Patent berdasarkan Negara

Proses pengidentifikasian tema pada judul patent yang berhasil di unduh dilakukan dengan menggunakan algoritma Bisection K-means pada spark, langkah awal dari menetapkan jumlah topik awal yang selanjutnya akan diproses lebih lanjut dengan Algoritma Bisection K-means. Jumlah

topik ditentukan berdasarkan plot antara banyaknya topik dengan nilai *sum square error* (SSE) seperti ditunjukkan pada Gambar 3.



Gambar 3 Jumlah Tema data Patent

Pada Gambar 3 didapat informasi awal berkenaan dengan jumlah *cluster* yang akan menjadi input dari algoritma Bisection K-means sebesar 18 cluster. Hasil anaisis cluster dengan input awal sebasr 18 cluster menhgasilkan cluster degan tema seperti ditunjukkan pada Gambar 4.

```
## cluster 1:  Group.1 agents cell novel uses
## cluster 2:  Group.1 disease cancer disorders pharmaceutical
## cluster 3:  Group.1 apparatus derivatives comprising composition
## cluster 4:  Group.1 antibodies compounds use novel
## cluster 5:  Group.1 inhibitors using systems kinase
## cluster 6:  Group.1 thereof use method composition
## cluster 7:  Group.1 uses thereof acid compounds
## cluster 8:  Group.1 method system disease detection
## cluster 9:  Group.1 disease treatment method inflammatory
## cluster 10:  Group.1 compositions methods treatment disease
## cluster 11:  Group.1 using methods treating disease
## cluster 12:  Group.1 methods related diagnosis alzheimers
## cluster 13:  Group.1 treating disorders diseases methods
## cluster 14:  Group.1 use thereof compositions methods
## cluster 15:  Group.1 treatment disease prevention use
## cluster 16:  Group.1 diseases treatment metabolic neurodegenerative
## cluster 17:  Group.1 comprising composition prevention treatment
## cluster 18:  Group.1 preventing composition treating pharmaceutical
```

Gambar 4 *Term Cluster*

4. Kesimpulan

Metode pengklasteran *Bisection K-means* mengelompokkan data twitter dengan kata kunci *disease* menghasilkan 18 buah klaster, Metode *Bisection K-means* yang diimplementasikan pada R masih terbatas dalam hal jumlah dimensi dari *term* yang terbentuk serta jumlah term yang dapat diretrieve dari database *espacenet*, Keterbatasan lainnya bahwa metode yang ada saat ini hanya berlaku untuk bahasa inggris, untuk dapat diterapkan pada bahasa Indonesia diperlukan algoritma dan database kata dasar dalam bahasa Indonesia. Sehingga hal ini menjadi saran bagi peneliti lainnya untuk dapat mengimplementasikan metode ini dalam bahasa Indonesia.

Ucapan Terima Kasih

Terima kasih kami ucapkan kepada Ketua Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Padjadjaran dan Staf Fakultas matematika dan Ilmu Pengetahuan Alam Universitas Universitas Padjadjaran yang telah memberikan dana Hibah penelitian sehingga penelitian ini dapat berjalan dengan lancar

Daftar Pustaka

Armbrust,M., Xin, R.S., Lian,C., Huai, Y. Spark SQL:Relational data processing in Spark. In SIGMOD, pages 1383–1394, 2015.

Barnett, M., Chandramouli,B., DeLine,R., Drucker,S., Fisher,D., Goldstein,J., Morrison,P., and Platt.J. Stat!: An interactive analytics environment for big data. In SIGMOD 2013, pages 1013–1016.

Das,S., Sismanis,Y., Beyer,K.S., Gemulla,R., P. Haas,J., and McPherson,J., Ricardo: integrating R and Hadoop. In SIGMOD 2010, pages 987–998. ACM, 2010.

Dou Henri JM (2004). Benchmarking R&D and companies through patent analysis using free databases and special software: a tool to improve innovative thinking. World Patent Information 26, 297–309.

Han, J., and Kamber, M. 2006. *Data Mining Concept and Techniques.*Morgan Kaufman Publisher. United States of America.

Kornacker,M., Behm,A., Bittorf,V., Bobrovytsky,T.,C. Ching,C., Choi,A., Erickson, J.,Grund,M., Hecht,D., Jacobs,M., Impala: A modern, open-source SQL engine for Hadoop. In CIDR 2015.

Gonzalez,J.E., R. Xin,S., Dave,A., Crankshaw,D., Franklin,J.M. and Stoica,I. Graphx: Graph processing in a distributed dataflow framework. In OSDI 2014, pages 599–613.

Meng, X.,Bradley,J.K.,Yavuz,B., Sparks, MLlib: Machine Learning in Apache Spark. CoRR, abs/1505.06807,2015

Prasad,S., Fard,S., Gupta,V., Martinez,J., LeFevre,J., Xu,V. Hsu,M., and Roy,I., Large-scale predictive analytics in vertica: Fast data transfer distributed model creation, and in-database prediction. In SIGMOD 2015.

Rostaing H., La bibliométrie et ses techniques. (Toulouse, Sciences de la Société, 1996).

Singh, Shalini S., et al. 2011. *K-means v/s K-medoids: A Comparative Study (National Conference on Recent Trends in Engineering and Technology)*

Sparks, Evan; Talwalkar, Ameet (2013-08-06). *"Spark Meetup: MLbase, Distributed Machine Learning with Spark"*. slideshare.net. Spark User Meetup, San Francisco, California. *Retrieved 10 February 2014.*

Steinbach M., Karypis, G. and Kumar. V. *"A comparison of document clustering*

ISBN 978-602-72216-1-1

techniques", Workshop on Text Mining, KDD, 2000

Venkataraman,S., Bodzsar,E., Roy,I., AuYoung, A. and Schreiber,S.R., Presto: Distributed Machine Learning and Graph Processing with Sparse Matrices. In Eurosys 2013, pages 197–210.

Wickham,H. and Francois,R. dplyr: A Grammar of Data Manipulation, 2015. R package version 0.4.3.

Zhang,Y., Zhang,W., and Yang.J., I/O-efficient statistical computing with RIOT. In ICDE 2010, pages 1157–1160.