


# Big Data: Menuju Evolusi Era Informasi Selanjutnya

Hapnes Toba

## Related papers

[Download a PDF Pack](#) of the best related papers 



[Simulasi dan komunikasi digital](#)  
Fikra Ihdina

[buku simulasi dan komunikasi digital bagian 2.pdf](#)  
ARDI JUNAIDI

[Big ; Data, Data Analyst, and Improving the Competence of Librarian](#)  
Ahmad Fauzi

# Big Data: Menuju Evolusi Era Informasi Selanjutnya

Hapnes Toba

Fakultas Teknologi Informasi  
Universitas Kristen Maranatha  
Jl. Suria Sumantri 65 Bandung 40164  
hapnes.toba@itmaranatha.org

**Abstract** — Dalam makalah<sup>1</sup> ini dipaparkan suatu kajian mengenai peran *big data* dalam riset-riset teknologi informasi terkini, pengaruh, serta potensinya dalam perkembangan ilmu komputer. Secara khusus dipaparkan mengenai peran temu balik informasi sebagai salah satu bidang riset dalam ilmu komputer untuk pengelolaan *big data*. Dari beberapa eksperimen yang telah dilakukan, dengan menggunakan pendekatan dari bidang ilmu temu balik informasi tersebut, dapat disampaikan empat isu besar yang menyertai perkembangan dan pengelolaan *big data*, yaitu: 1). infrastruktur jaringan komputer; 2). normalisasi (penyamaan) persepsi terhadap data (*data science*); 3). *big data* sebagai fondasi untuk *collaborative intelligence*; dan 4). perlunya kesadaran akan nilai informasi yang mengiringi citra diri seseorang, sebuah organisasi ataupun institusi.

**Keywords** — *big data*, temu balik informasi, penyaringan data, *data science*.

## I. KONSEP BIG DATA

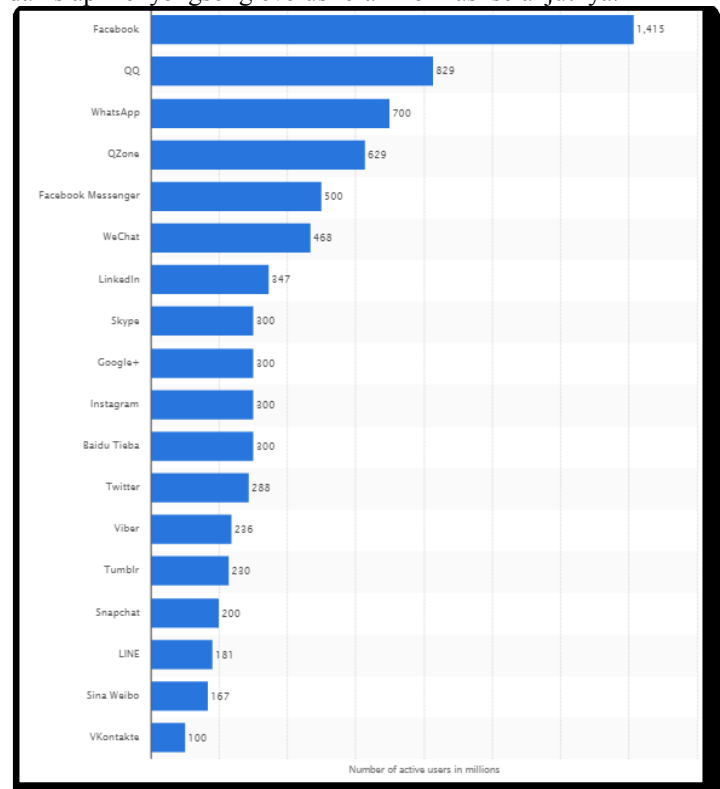
Beranjak dari tema umum yang ditawarkan dalam SeTISI 2015, yaitu: 'Peran keamanan informasi menuju Indonesia hebat dalam menghadapi *Asean Economic Community* 2015', maka dalam makalah ini dilakukan kajian mengenai suatu konsep, yaitu *big data*. Konsep ini mulai muncul sebagai tren dalam pengelolaan informasi dalam kurun lima tahun terakhir mengingat begitu besarnya pertumbuhan data di Internet, khususnya melalui media sosial [1, 2, 3].

Secara umum *big data* dapat diartikan sebagai sebuah kumpulan data yang berukuran sangat besar (*volume*), sangat cepat berubah/bertumbuh (*velocity*), hadir dalam beragam bentuk/format (*variety*), serta memiliki nilai tertentu (*value*), dengan catatan jika berasal dari sumber yang akurat (*veracity*) [1, 3]. Hal utama yang membedakan *big data* dengan kumpulan data konvensional terletak pada mekanisme pengelolaannya. Sistem basis data relasional yang saat ini umum digunakan, sudah dirasakan tidak mampu menangani kompleksitas *big data* secara optimal [4].

Dengan disadari atau tidak, masyarakat moderen saat ini, khususnya di kota-kota besar, termasuk di Indonesia telah memiliki ketergantungan terhadap Internet. Sebagai suatu infrastruktur, kehadiran Internet telah bertumbuh menjadi kebutuhan untuk berkomunikasi, bertukar pikiran, bahkan menjadi suatu saluran untuk mencurahkan isi hati.

Sebuah implikasi langsung dari kebutuhan akan Internet tersebut adalah pertumbuhan data secara masif yang memenuhi simpul-simpul jaringan (*servers*) seantero jagad. Salah satu hal yang mungkin tidak disadari adalah: dengan adanya ketersebaran data tersebut menyebabkan 'kemudahan' dalam mengakses informasi – yang baik ataupun buruk – tentang individu, organisasi ataupun institusi.

Oleh karena itulah dalam makalah ini dipaparkan suatu kajian khusus mengenai isu-isu seputar pengelolaan *big data*. Diharapkan bahwa kajian dalam makalah ini dapat memberikan suatu kesadaran (*awareness*), sehingga sebagai komunitas ilmiah, kehadiran Internet, mesin temu balik dan media sosial dapat dimanfaatkan secara positif untuk mendukung citra bangsa Indonesia yang lebih hebat dan siap menyongsong evolusi era informasi selanjutnya.



Gambar 1. Statistik pengguna media sosial Maret 2015<sup>2</sup>.

<sup>1</sup> Dibawakan dalam sesi pembicara tamu (*keynote speaker*) pada SeTISI 9 April 2015 di UK. Maranatha Bandung.

<sup>2</sup> <http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> (akses April 2015)

## II. PENGELOLAAN *BIG DATA*

### A. Sumber Perolehan Data

Dilihat melalui kacamata riset-riset ilmiah terkini, sumber-sumber perolehan *big data* dapat dikelompokkan ke dalam tiga sumber utama, yaitu: media sosial dan *blogs*, lalu lintas data (secara eksternal/internal), dan mesin temu balik informasi (*search engine*) [1, 2, 5]. Media sosial memiliki kontribusi yang sangat besar sebagai sumber data 'tersembunyi' yang sangat cepat berubah dan hadir dalam berbagai formatnya, baik berupa teks, gambar maupun video (*multimedia*). Sebagai ilustrasi, jumlah besaran data yang dihasilkan di dunia maya secara keseluruhan dalam tahun 2011, diprediksi mencapai 1.8 Zetta Byte ( $\approx 10^{21}$ B), dan jumlah tersebut akan naik dua kali lipat setiap dua tahun [1, 14]. Prediksi kenaikan tersebut juga diperkuat dengan statistik pengguna media sosial aktif pada bulan Maret 2015 dalam Gambar 1, yang akan terus bertambah.

Selain bersumber pada media sosial, lalu lintas data dalam suatu jaringan komputer (*enterprise data*) dapat pula dianggap sebagai *big data*. Ukuran yang dapat dianggap sebagai *big data* adalah jika dalam jaringan tersebut mengalir data pada kisaran Terra sampai Peta Byte per hari, dalam berbagai bentuknya, seperti: akses aplikasi, *sharing file*, email, *chatting*, ataupun aktivitas jaringan lainnya.

Sumber lain yang dapat dianggap pula sebagai *big data* adalah Internet *search engine* (SE). Meskipun SE tidak secara langsung menghasilkan materi data yang diinginkan, namun SE merupakan sebuah pintu gerbang yang menjadi mediator sumber data sesungguhnya, misalnya dalam konteks *ad-hoc search* [6, 7]. Lebih jauh lagi, melalui kehadiran SE sumber-sumber penyedia *big data* yang 'terpercaya' dapat diakumulasi dan mempermudah penyaringan informasi yang diperlukan. Sebagai contoh pada saat flu burung mewabah di tahun 2010, lalu lintas data pada SE Google mengalami peningkatan drastis dengan kueri-kueri yang sangat spesifik tentang penyakit tersebut [1, 6].

### B. Penyimpanan dan Akses Data

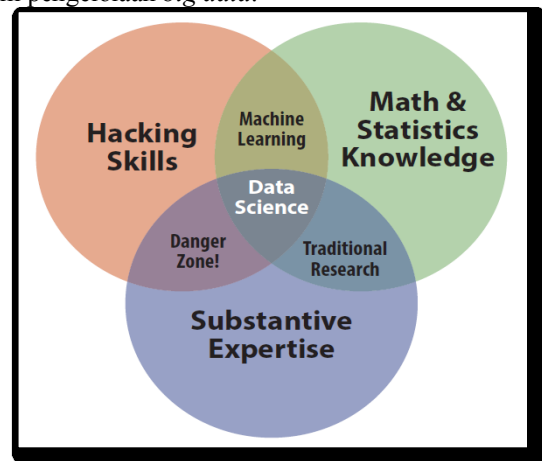
Mengacu pada definisi *big data* pada bagian terdahulu, diperlukan suatu penanganan khusus terhadap konten *big data*, ditinjau secara ukuran dan bentuknya. Konsep pengelolaan data secara relasional yang berbasis pada data-data terstruktur, dirasakan tidak mumpuni sebagai media penyimpanan *big data*, yang pada umumnya sangat tidak terstruktur dan terdistribusi. Server-server raksasa seperti pada perusahaan Google, Yahoo!, ataupun Facebook menggunakan teknologi *Storage Area Network* (SAN) yang dilengkapi dengan kemampuan *cloud* dan *grid computing*, misalnya dengan menggunakan Hadoop [8, 9].

Basis data berbasis pada NoSQL menjadi salah satu alternatif yang saat dianggap cocok untuk menangani *big data* [10]. Alternatif lain yang juga populer adalah dengan menggunakan basis data berbasis dokumen, seperti MongoDB yang menyimpan dokumen sebagai *binary JSON* (BSON) [11].

Selain teknologi penyimpanan dalam basis data, *big data* juga memerlukan teknik pemrosesan dan pemrograman yang berbeda dibandingkan dengan basis data relasional. Pemrograman dengan MapReduce [12, 13], misalnya telah menjadi suatu standar untuk pemrosesan data secara paralel dan terdistribusi dalam berbagai *server*. Guna meningkatkan efisiensi pemrograman, perusahaan-perusahaan besar mengembangkan berbagai bahasa pemrograman yang berbasis pada MapReduce, misalnya: Sawzall (Google) [15], PigLatin (Yahoo!) [16], Hive (Facebook) [17], dan Scope (Microsoft) [18].

### C. Aplikasi dan Pemanfaatan

Dengan mengingat begitu luasnya sumber yang tersedia, jika *big data* hanya dipandang sebagai suatu kumpulan data saja, maka tidak akan menghasilkan informasi apapun. Untuk dapat menghasilkan informasi dari *big data*, diperlukan adanya analisis yang mendalam, misalnya melalui pemodelan matematis. Gambar 2 memperlihatkan suatu ilustrasi bagaimana *data science*, dapat berperan dalam pengelolaan *big data*.



Gambar 2. *Data science* dalam konteks *Big Data*<sup>3</sup>.

- **Hacking skills:** dalam konteks ini jangan diartikan sebagai suatu bentuk negatif, namun harus dibawa pada ranah 'manipulasi' data secara positif. Termasuk di dalam konteks ini adalah kemampuan untuk menyamakan persepsi tentang berbagai format data (*pre-processing/normalization/discretization*). Misalnya: bagaimana mengaitkan antara informasi sebuah *blog* rumah sakit dan penanganan suatu penyakit, sehingga calon pasien memiliki gambaran yang lebih lengkap tentang penanganan penyakitnya di rumah sakit tersebut [19].
- **Substantive expertise:** dalam konteks ini diperlukan cara pandang dari seorang ahli untuk dapat menginterpretasikan data yang diperoleh (*feature engineering*). Misalnya: bagaimana mengartikan kalimat-kalimat yang dituliskan seseorang dalam

<sup>3</sup> <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram> (akses April 2015)

sebuah email/posting dalam media sosial sehingga dapat dijadikan sebagai model untuk menebak kepribadian seseorang [20, 21, 22].

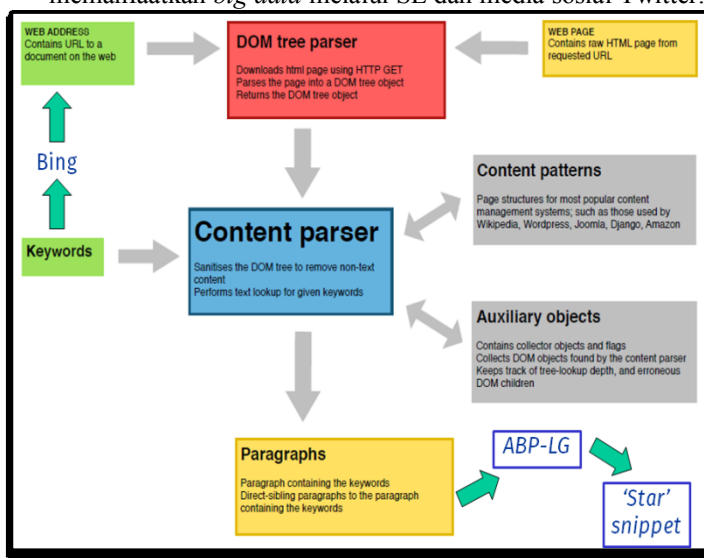
Selain diperlukan pengetahuan tentang konten data itu sendiri, di sisi lain diperlukan pula mekanisme untuk menelaah apakah data yang terkumpul adalah data yang valid dan sah (tidak melanggar privasi). Dalam kaitan ini, SE besar seperti Google telah memiliki *disclaimer*<sup>4</sup> yang menyatakan bahwa semua yang dikumpulkan tentang seseorang/institusi adalah hal yang telah 'disetujui' [23].

Hal ini tanpa sadar sering kali diabaikan, misalnya: pada saat proses registrasi akun email, pada saat mengkoneksikan telepon pintar dengan akun Google, ataupun penggunaan *cookies* pada *web browser*. Dengan demikian pada saat informasi terkumpul melalui SE, sebagian dari isu privasi tersebut sudah tereduksi.

- **Math and Statistics Knowledge:** pengetahuan tentang pemodelan matematis yang cocok untuk sebuah problem tertentu sangat diperlukan dalam pemanfaatan *big data*. Pemodelan umum yang banyak dipakai dalam pembelajaran mesin [1, 3, 24], seperti: pembentukan kluster, *variable* tersembunyi (*hidden/latent variables*), analisis korelasi, *graphical models*, ataupun teknik klasifikasi, masih sangat diperlukan dalam mencari kaitan antar kumpulan data.

### III. TEMU BALIK INFORMASI DALAM *BIG DATA*

Dalam bagian ini akan diuraikan contoh aplikasi yang memanfaatkan *big data* melalui SE dan media sosial Twitter.



Gambar 3. Proses penemuan jawaban dengan SE Bing.

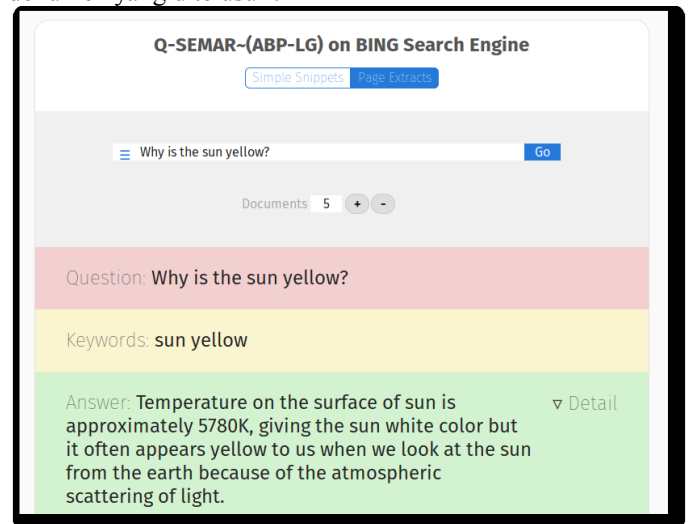
#### A. Sistem Tanya Jawab

Sebuah sistem tanya jawab (STJ) memiliki tugas utama mengembalikan sebuah jawaban terhadap pertanyaan yang diajukan dalam bahasa natural. STJ dapat berbasis pada

komunitas (*users generated content*) seperti dalam Yahoo!Answers ataupun sistem mandiri. Forum-forum evaluasi STJ seperti TREC ataupun CLEF telah membuat teknologi STJ turut dimanfaatkan dalam SE [25].

Salah satu sumber data yang dapat digunakan dalam penemuan jawaban adalah hasil SE. Dengan menelusuri hasil SE dan dengan model penemuan jawaban berbasis struktur kalimat, dapat ditentukan bagian teks tertentu (berupa frasa) yang memberikan kemungkinan jawaban terbesar [26]. Sebagai contoh dalam Gambar 3, STJ memanfaatkan SE Bing, dan menghasilkan jawaban pada Gambar 4, untuk pertanyaan: 'Why is the sun yellow?'.

Hasil evaluasi dalam riset menunjukkan bahwa dengan memanfaatkan hasil SE diperoleh tingkat akurasi yang lebih tinggi dibandingkan dengan koleksi dokumen statis. Hal ini menunjukkan bahwa pengolahan *big data* sangat menjanjikan dalam temu balik informasi, khususnya untuk penemuan jawaban. Ditinjau secara arsitektural, STJ yang diusulkan sangat sederhana dalam Gambar 3, namun memerlukan daya pemrosesan yang cukup besar untuk melakukan ekstraksi fitur dan proyeksi model jawaban. Selain itu diperlukan koneksi Internet yang stabil untuk menjamin diperolehnya teks yang representatif dari dokumen yang ditelusuri.



Gambar 4. Contoh pertanyaan dengan memanfaatkan isi halaman web.

#### B. Prediksi Profesi Menggunakan Twitter

Salah satu riset yang menjadi tren saat ini adalah pembentukan model untuk prediksi profesi melalui kata-kata yang muncul dalam kumpulan data *microblog* (Twitter). Sasaran dari penelitian ini adalah untuk memanfaatkan kehadiran kata-kata, yang diasumsikan sebagai gaya bahasa, sebagai cara untuk mendeteksi suatu profesi [27] atau menentukan sentimen tentang seseorang atau institusi [28].

Sebuah riset yang saat ini sedang dilakukan di Fakultas Teknologi UK. Maranatha adalah melakukan deteksi profesi melalui gaya bahasa dari sebuah *tweet*. Data-data diambil dari Twitter antara bulan Juni-Oktobre 2014, sejumlah sekitar 10 ribu *tweets* untuk 4 jenis pekerjaan, yaitu: politikus, pelajar, artis dan musisi. Dalam masa tersebut di

<sup>4</sup> <http://www.google.com/policies/> (akses April 2015)

Indonesia sedang ramai proses pemilihan presiden dan polemik hasil pemilu. Dari hasil analisis awal diketahui bahwa para pemberi *tweet* tersebut didominasi oleh para politikus, kaum remaja (pelajar) dan para artis / musisi yang dipakai partai tertentu untuk menarik pendukung.

Gambar 5 memberikan contoh hasil pembentukan bobot kata pada setiap model profesi untuk data yang terkumpul, dengan pembuangan *stopwords* untuk bahasa Indonesia. Model Bayes terbaik yang diusulkan melalui penelitian ini adalah dengan metode multinomial melalui pemecahan kata secara *unigram* melalui penghilangan *stopwords*, dengan rata-rata akurasi 92.44% untuk keempat jenis profesi yang diuji.

Hasil evaluasi menunjukkan bahwa data dalam media Twitter:

1. Sangat dinamis: kata-kata penting berubah dari waktu ke waktu).
2. Banyak terjadi percampuran bahasa terutama bahasa Inggris dan Indonesia.
3. Banyak dipakainya kata-kata informal, seperti singkatan, bahasa *slang* ataupun bahasa makian, yang sangat berpengaruh dalam menentukan sentimen terhadap suatu *tweet*.

Temuan-temuan di atas mengindikasikan pula bahwa untuk mengolah *big data* diperlukan mekanisme *pre-processing* yang lebih kompleks dibandingkan teks atau data terstruktur pada umumnya.

<b>Nilai bobot politisi:</b> SBY 7.846214E-4 Ketum 7.846214E-4 perlu 8.282115E-4 pemerintah 8.282115E-4 rakyat 8.282115E-4 kota 9.1539166E-4 punya 9.1539166E-4 warga 9.5898175E-4 ekonomi 0.0011769321 Indonesia 0.0013512925 s 0.19489124	<b>Nilai bobot artis:</b> is 0.0034364262 I 0.0034835003 of 0.003907169 @ 0.0040483926 by 0.004095467 at 0.004424987 with 0.0046132845 a 0.0046603587 you 0.004707433 my 0.005131102 and 0.0054606223 s 0.09885609
<b>Nilai bobot pelajar:</b> I 0.0047124447 in 0.0054337373 of 0.0057703406 is 0.0057703406 for 0.006587805 My 0.006732064 at 0.0068282364 you 0.0077899597 (at 0.007886132 a 0.008463166 to 0.008944028 and 0.009953837 The 0.010915561 s 0.080929026	<b>Nilai bobot musisi:</b> musik 9.349766E-4 nyanyi 9.774755E-4 jazz 0.0010199745 tiket 0.0011899703 baru 0.0011899703 band 0.0012324692 album 0.0018699532 konser 0.002039949 nonton 0.002039949 the 0.009094773 s 0.10773481

Gambar 5. Pembobotan kata-kata penting dalam berbagai model 'profesi'

### C. Web People Act

Jenis riset ini dilakukan di Fakultas Teknologi UK. Maranatha, sebagai suatu perluasan dari riset *Web People Search* (WePS). Tugas utama dalam WePS adalah untuk membedakan (*disambiguation*) nama-nama orang tertentu berdasarkan informasi dasar yang diperoleh dari SE [29].

Dalam *Web People Act*<sup>5</sup> (WePA), beberapa informasi dasar telah dimiliki, misalnya terkait alumni, yaitu: nama universitas dan nama fakultasnya. Namun informasi dasar tersebut ingin dilengkapi dengan jenis informasi personal lainnya, seperti: perubahan alamat email atau nomor telepon, media sosial yang diikuti dalam dunia maya, dan prediksi jenis profesi yang digeluti.

Salah satu teknik dari temu balik informasi yang dapat digunakan adalah dengan melakukan ekspansi kueri berdasarkan frekuensi kata-kata penting dari hasil SE [30, 31]. Dengan melakukan ekspansi kueri diharapkan bahwa terjadi penyaringan informasi yang lebih spesifik terkait data diri seseorang.

Selain menggunakan frekuensi kata-kata penting, dapat diusulkan penyaringan informasi dengan mengambil tautan-tautan berpengaruh (*influential links*). Hal ini dilakukan dengan teknik *suffix stripping* pada URL, sehingga menghasilkan sekumpulan nama entitas (*links*) untuk diproyeksikan pada situs-situ populer<sup>6</sup> di dunia. Dengan cara ini akan dapat diketahui aktivitas seseorang, apakah dia aktif dalam sosial media atau halaman web lainnya. Gambar 6 dan 7 memberikan contoh jalannya eksekusi WePA untuk ekspansi kueri dasar dan penyaringan tautan.

Gambar 6. Masukan informasi WePA: nama dan informasi dasar.

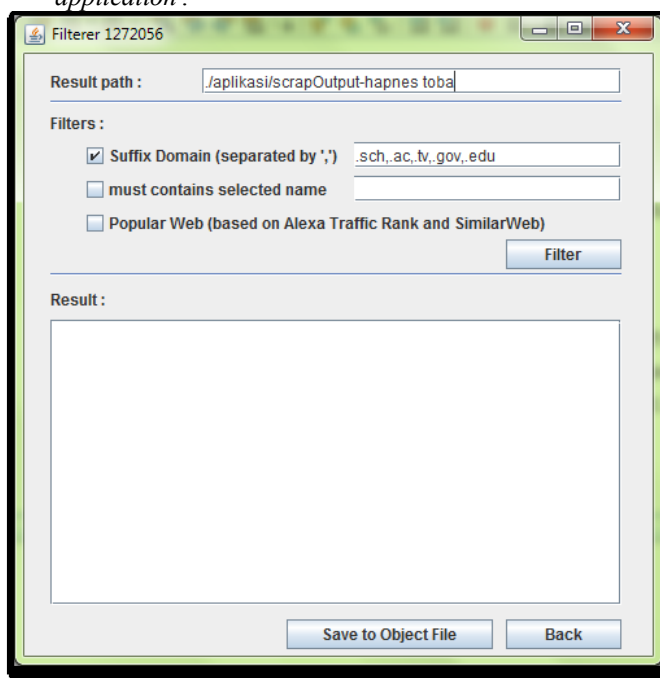
<sup>5</sup> Pada saat makalah ini dituliskan, riset masih berjalan.

<sup>6</sup> <http://www.alex.com/> (akses April 2015)



Dari eksperimen awal yang telah dilakukan, dapat ditarik kesimpulan sementara sebagai berikut:

1. Aktivitas seseorang di dunia maya besar kemungkinannya untuk dapat ditebak berdasarkan frekuensi kata-kata tertentu yang mengikat diri orang tersebut. Entah itu berupa kata-kata yang ditulisnya sendiri atau komentar dari pihak lain.
2. Dengan memanfaatkan hasil SE dapat diketahui aktivitas seseorang dalam dunia maya, dan dapat terlihat pula halaman web apa saja yang aktif dikunjungi atau memuat informasi tentang orang tersebut.
3. Untuk dapat menebak dengan lebih tepat aktivitas seseorang, akan diuji coba lebih jauh menggunakan proyeksi kata-kata pada suatu kamus – misalnya dengan WordNet<sup>7</sup> – untuk mencari tautan meronim, hipernim, hiponim, ataupun sinonim dari kemunculan kata tertentu, dan divisualisasikan pada suatu graf, misalnya melalui WordVis<sup>8</sup>, pada Gambar 8, untuk sebuah kata 'application'.



Gambar 7. Penyaringan tautan berdasarkan popularitas.

#### IV. KESIMPULAN DAN POTENSI RISET

Mengacu kepada kajian mengenai *big data* dan beberapa riset yang telah atau sedang dilakukan, beberapa hal yang dapat disampaikan sebagai kesimpulan adalah:

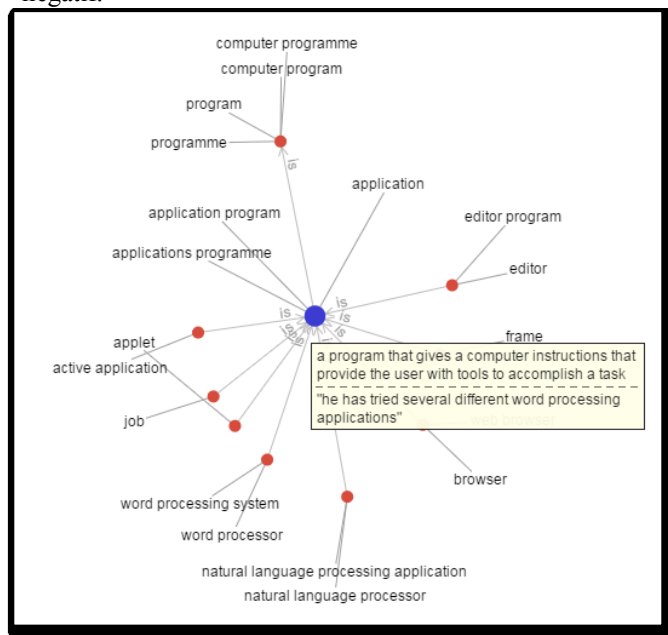
1. Pemanfaatan *big data* sangat memerlukan infrastruktur Internet yang mumpuni sehingga dapat menjamin bahwa data-data dalam jumlah besar dapat diakses secara kontinu.
2. Dengan begitu banyak dan bervariasi data yang 'tersedia' (baik disadari maupun tidak disadari) secara

<sup>7</sup> <https://wordnet.princeton.edu/> (akses April 2015)

<sup>8</sup> <http://wordvis.com> (akses April 2015)

*online*, menyebabkan dimungkinkannya pengolahan informasi yang tepat guna sesuai kebutuhan. Untuk hal ini tentu saja diperlukan adanya pemodelan data (*data science*) yang mumpuni.

3. Dengan tersedianya berbagai komunitas yang aktif di dunia maya, sangat dimungkinkan adanya kolaborasi antara berbagai bidang ilmu.
4. Perlunya menumbuhkan kesadaran akan nilai data yang di-*posting* dalam dunia maya. Meskipun data sangat tersebar (*sparse*) dalam berbagai sumber, namun dengan kehadiran SE menjadi sangat mudah diperoleh. Melalui teknik pengolahan informasi yang 'tepat' data-data tersebut dapat menjadi hal yang positif ataupun negatif.



Gambar 8. Contoh visualisasi kata 'application' dengan WordVis yang berbasis pada WordNet.

Adapun beberapa potensi riset yang dalam waktu dekat ini dapat diusulkan mencakup hal-hal sebagai berikut:

1. Menghimpun dan membentuk *dataset* standar sebagai sarana evaluasi untuk pemodelan yang lebih tepat guna sebelum diterapkan untuk permasalahan ataupun pada data yang lebih besar. Suatu contoh *dataset* standar dapat dilihat pada [32].
2. Memperdalam teknik untuk memvisualisasi hasil pemodelan *big data*, sehingga dapat menjadi sarana pendukung untuk pengambilan keputusan, terutama dalam organisasi.
3. Membentuk mekanisme pengamanan, otorisasi dan privasi data, terutama untuk data-data yang disebarkan melalui media *online*, khususnya bagi institusi, seperti perguruan tinggi, yang tentu saja memiliki kepentingan pencitraan dalam masyarakat.
4. Memperlengkapi kurikulum pendidikan ilmu komputer di perguruan tinggi dengan materi pengolahan *big data*, baik dari sisi *soft skills*, penyediaan maupun pengelolaan data. Hal ini diperlukan sebagai salah satu

persiapan guna menyongsong evolusi era informasi selanjutnya.

#### DAFTAR PUSTAKA

- [1] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2), 171-209.
- [2] Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.
- [3] Jou, S. (2014). Towards a Big Data Behavioral Analytics Platform. <http://interaset.com>. Access April 2015.
- [4] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., ... & McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.
- [5] Crawford, K. (2011). Six provocations for big data. [http://papers.ssrn.com/sol3/Papers.cfm?abstract\\_id=1926431](http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1926431). Access April 2015.
- [6] Lazer, D. M., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. [http://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20\(WP-Final\).pdf?sequence=1](http://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20(WP-Final).pdf?sequence=1). Access April 2015.
- [7] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, 36(4), 1165-1188.
- [8] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.
- [9] Luo, T., Chen, G., & Zhang, Y. (2013). H-DB: Yet Another Big Data Hybrid System of Hadoop and DBMS. In *Algorithms and Architectures for Parallel Processing* (pp. 324-335). Springer International Publishing.
- [10] Moniruzzaman, A. B. M., & Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*.
- [11] Chodorow, K. (2013). *MongoDB: the definitive guide*. " O'Reilly Media, Inc."
- [12] Chen, Y., Alspaugh, S., & Katz, R. (2012). Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads. *Proceedings of the VLDB Endowment*, 5(12), 1802-1813.
- [13] Madden, S. (2012). From databases to big data. *IEEE Internet Computing*, 16(3), 4-6.
- [14] Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iView*, (1142), 9-10.
- [15] Criou, N., Mahadevan, C., & Venkatakrishnan, S. (2013). *U.S. Patent No. 8,463,830*. Washington, DC: U.S. Patent and Trademark Office.
- [16] Olston, C., Elmeleegy, K., & Reed, B. (2013). *U.S. Patent No. 8,356,050*. Washington, DC: U.S. Patent and Trademark Office.
- [17] Menon, A. (2012, September). Big data@ facebook. In *Proceedings of the 2012 workshop on Management of big data systems* (pp. 31-32). ACM.
- [18] Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. *interactions*, 19(3), 50-59.
- [19] Grajales III, F. J., Sheps, S., Ho, K., Novak-Lauscher, H., & Eysenbach, G. (2014). Social media: a review and tutorial of applications in medicine and health care. *Journal of medical Internet research*, 16(2), e13.
- [20] Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., ... & Seligman, M. E. (2014). The online social self an open vocabulary approach to personality. *Assessment*, 21(2), 158-169.
- [21] Gou, L., Zhou, M. X., & Yang, H. (2014, April). KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 955-964). ACM.
- [22] Gritzalis, D., Kandias, M., Stavrou, V., & Mitrou, L. (2014). History of information: The case of privacy and security in social media. In *Proc. of the History of Information Conference* (pp. 283-310).
- [23] ... Google Privacy Policy. (2015). [http://static.googleusercontent.com/media/www.google.com/en/intl/en/policies/privacy/google\\_privacy\\_policy\\_en.pdf](http://static.googleusercontent.com/media/www.google.com/en/intl/en/policies/privacy/google_privacy_policy_en.pdf). Access April 2015.
- [24] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on*, 26(1), 97-107.
- [25] Peñas, A., Magnini, B., Forner, P., Sutcliffe, R., Rodrigo, Á., & Giampiccolo, D. (2012). Question answering at the cross-language evaluation forum 2003–2010. *Language resources and evaluation*, 46(2), 177-217.
- [26] Toba, H. (2015). *Pemodelan Frasa Pengandung Jawaban (ABP-LG) Untuk Sistem Tanya Jawab* (Doctoral dissertation). Faculty of Computer Science, Universitas Indonesia. (Dis-38 (Softcopy Dis-29) Source code Dis-17).
- [27] Ramírez-de-la-Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H., & Sánchez-Sánchez, C. (2014). Towards Automatic Detection of User Influence in Twitter by Means of Stylistic and Behavioral Features. In *Human-Inspired Computing and Its Applications* (pp. 245-256). Springer International Publishing.
- [28] Sarlan, A., Nadam, C., & Basri, S. (2014, November). Twitter sentiment analysis. In *Information Technology and Multimedia (ICIMU), 2014 International Conference on* (pp. 212-216). IEEE.
- [29] Berendsen, R., Kovachev, B., Nastou, E. P., de Rijke, M., & Weerkamp, W. (2012). Result disambiguation in web people search. In *Advances in Information Retrieval* (pp. 146-157). Springer Berlin Heidelberg.
- [30] Mishne, G., Dalton, J., Li, Z., Sharma, A., & Lin, J. (2013, June). Fast data in the era of big data: Twitter's real-time related query suggestion architecture. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 1147-1158). ACM.
- [31] Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1), 1.
- [32] Wang, L., Zhan, J., Luo, C., Zhu, Y., Yang, Q., He, Y., ... & Qiu, B. (2014, February). Bigdatabench: a big data benchmark suite from internet services. In *High Performance Computer Architecture (HPCA), 2014 IEEE 20th International Symposium on* (pp. 488-499). IEEE.