Web Usage Mining untuk Mengetahui Pola Akses Pengguna Menggunakan Algoritma *Improved K-means* (Studi Kasus : Kitabisa.com)

Muthia Bianda, Eko Darwiyanto, M.T, dan **Gede Agung Ary Wisudiawan, S.Kom, M.T**

Abstrak : Kegiatan donasi yang pada awalnya dilakukan secara *offline,* sekarang bisa dilakukan melalui internet. Untuk mendukung donasi *online* ini, maka akan dicari aspek yang bisa diperbaiki dan dikembangkan dari sistem donasi, supaya penggunanya semakin bertambah dan nyaman menggunakan sistem. Salah satu perusahaan yang menyediakan layanan donasi adalah web kitabisa.com. Perbaikan website bisa dimulai dengan melihat pola akses pengguna website tersebut. Untuk melihat pola akses pengguna, digunakan metode *Web Usage Mining*. Pada proses pengelompokkan jenis pengguna berdasarkan pola aksesnya, digunakan algoritma K-Means. Namun algoritma *K-Means* mempunyai kendala untuk menentukan nilai K yang optimal [7]. Untuk menutupi kendala itu, maka digunakan algoritma *Improved K-Means* [2], yaitu mendapatkan nilai K dari proses algoritma perulangan *dataset*. Hasil dari penelitian ini adalah mengetahui nilai K yang paling optimal serta memberi informasi pengelompokan pengguna berdasarkan pola aksesnya.

Kata kunci : *Web Usage Mining*, *Server log*, pola akses pengguna website, *clustering*, *K-Means*, *Improved K-Means.*

1. Pendahuluan

Perkembangan zaman membawa aktifitas sehari-hari menjadi digital dan instan dengan bantuan teknologi Internet. Internet menjadi tempat untuk mencari dan penyebaran informasi, transaksi bisnis, sosial media, dan lain-lain. Salah satu aktifitas yang dibawa lebih praktis dengan internet adalah aktifitas donasi. Donasi atau penggalangan dana sampai saat ini merupakan komponen yang paling penting dan yang paling dibutuhkan untuk membantu korban bencana alam, membantu kegiatan sosial, kegiatan pendidikan, dll. Di Indonesia sendiri, ada tim Kitabisa yang membuat sistem penggalangan dana. Situs web kitabisa bisa diakses di http://www.kitabisa.com. Untuk mendukung donasi *online* ini, maka akan dicari aspek yang bisa diperbaiki atau dikembangkan dari sistemnya. Pencarian aspek tersebut bisa dimulai dari melihat pola akses pengguna di dalam website. Untuk melihat pola akses pengguna tersebut, maka akan diimplementasikan *Web Usage Mining*.

Web Usage Mining digunakan karena merupakan metode satu-satunya untuk mengetahui pola akses pengguna website. Dengan mengetahui pola pengguna, ada beberapa manfaat untuk website tersebut, misalnya untuk memperbaiki sistem dalam (*internal*) atau tampilan muka (*interface*) website. Secara umum, *Web Usage Mining* akan melakukan tiga proses besar berikut : *preprocessing, pattern discovery*, dan *pattern analysis*.

Di dalam tugas akhir ini, proses *clustering* yang berada di dalam tahap *pattern discovery*, akan menggunakan algoritma *K-Means*. Algoritma K-Means dipilih karena *K-Means* merupakan salah satu teknik *clustering* yang paling mudah, cepat, tidak memakan banyak memori, serta bisa diaplikasikan kepada *dataset* berukuran kecil atau besar [14]. Untuk menggunakan *K-Means*, nilai K harus ditentukan terlebih dahulu sebelum proses *clustering* dimulai. Nilai K umumnya ditentukan sendiri oleh pengguna algoritma ini, namun kekurangannya adalah pengguna tidak mengetahui apakah nilai K yang dipilih sudah yang paling tepat dari seluruh angka lainnya. Untuk menutupi kekurangan tersebut, maka digunakanlah algoritma *Improved K-Means*.

Improved K-Means [2] sudah terbukti bisa menghasilkan nilai K yang paling optimal. *Improved K-Means* mempunyai kelebihan untuk menjalankan algoritma *clustering* tanpa memilih nilai K terlebih dulu, karena hasil *clustering*nya sudah memberikan jumlah *cluster* yang paling optimal.

2. Tinjauan Pustaka

Di dalam bab ini, akan dijelaskan teori dan teknik apa saja yang digunakan.

2.1 Web Usage Mining

Web Usage Mining (WUM) adalah metode untuk mencari informasi pola akses pengguna situs web dengan memanfaatkan data web. Sumber data untuk melakukan WUM beragam, yaitu : *server level collection, proxy level collection,* dan *client level collection* [8]. Pada umumnya, yang digunakan adalah *server level collection* atau *server log*, karena *server log* menyimpan banyak *history browsing* dari banyak pengguna. *Client level collection* digunakan di level browser dan diimplementasikan dengan bantuan dari pihak ketiga, seperti Javascripts atau Java Applets. *Proxy level collection* juga bekerja di level browser. Proxy dapat digunakan untuk menyimpan alamat situs web agar situs web tersebut dapat diproses lebih cepat jika diakses kembali. *Client level collection* maupun *Proxy level collection* keduanya membutuhkan persetujuan dari pengguna web browser tersebut untuk berkerja sama. Di dalam tugas akhir ini, hanya akan menggunakan *server log*. Di dalam WUM, ada tiga proses utama, yaitu *Preprocessing, Pattern Discovery,* dan *Pattern Analysis.*

2.2 Sequence Based Analysis

Sequence Based Analytics adalah metode yang diperkenalkan oleh Park, Suresh, dan Jeong (2008), untuk melihat urutan aktifitas pengunjung di dalam sebuah website. Urutan aktifitas pengguna tersebut akan ditangkap oleh sebuah *Sequence Matrix*.

Algoritma untuk membuat sequence matrix adalah sebagai berikut :

Langkah 1. Inisiasi $M_s(i,j)$ dengan 0

Langkah 2. Insiasi $t = 0$

Langkah 3. $M_s(x_t, x_{t+1}) = M_s(x_t, x_{t+1}) + 1$

Langkah 4. $t = t + 1$

Langkah 5. Ulangi langkah 3 dan 4 jika *t* <= *h*

Langkah 6. Untuk setiap *i* dan *j*, $M_s(i,j)$ <- $M_s(i,j)$ / $\sum_{j=0}^{M} M_s(i,j)$

2.3 Improved K-means

Improved K-means adalah algoritma *clustering* baru yang dibuat sedemikian rupa supaya pengguna algorima *K-Means* tidak perlu menentukan nilai K. Proses ini memerlukan perulangan atau iterasi pembacaan *dataset,* dimana nilai K bertambah saat sebuah kondisi cocok.

Didalam algoritma ini diperlukan sebuah variabel bernama α yang bertugas sebagai nilai pembeda antar objek. Nilai α ditentukan sendiri oleh pengguna algortima ini. Adapun algoritmannya seperti berikut [2] :

Algoritma 1 memilih *initial points* :

1. Input : Dataset D yang mempunyai N objek.
2. Pilih satu object (X_1) dari dataset.
3. Jadikan X_1 sebagai *center* dari *cluster* pertama :
   K = 1, $B_1 = X_1$
   For *i* = 2 to N
   Cari B_m: $d(X_i, B_m) = \min_{1 \le k \le K} d(X_i, B_k)$
   Jika d(Xi, Bm) > ∝ maka k = k + 1; Bk = Xi
   atau i = i + 1
4. Output : B1, B2,…., Bk

Hasil dari Algoritma 1 adalah nilai K dan sudah ada *initial point* berjumlah K. Dari *initial point* tersebut, dilakukan algoritma *clustering K-Means* biasa (bab 2.4). Namun hasil dari algoritma 1 masih menghasilkan jumlah *cluster* yang banyak, sehingga hasil *clustering K-Means* tadi kemudian dinilai apakah ada *cluster* yang bisa disatukan. Proses ini dilakukan di algoritma 2. Di dalam algoritma 2, diperlukan satu nilai *threshold,* yaitu β min, yang

menetapkan apakah sebuah *cluster* bisa disatukan dengan *cluster* lainnya. Nilai β min ditetapkan sendiri oleh pengguna algoritma ini.

Algoritma 2

Input : Hasil *clustering* dari algoritma 1 dan *threshold* β min

Ouput : Hasil akhir *clustering*

Metode :

1. Hitung radius dari tiap *cluster*
2. Hitung jarak matriks antara semua *center cluster centerdis*
3. Hitung faktor matriks spsifik :
   $\beta = \beta(\ldots), (1 \leq \ldots \leq \ldots$
   3.1 Cari radius dari Ci dan Cj jika *centerdis*(i,j)<=(Ci+Cj) ke langkah 3.3 atau $\ldots = 0$
   3.2 Hitung objek di dalam Ci dan Cj
   3.3 Hitung $\ldots$ = common_object/total_object
4. Cari spesifik faktor terbesar $\ldots$ dan jika $\ldots >$ $\ldots$ makansatukan *cluster* Ci dan Cj, atau ke langkah 6.
5. Perbaharui spesifik faktor terbesar dan pergi ke langkah 4
6. Perharui semua *center cluster*.
7. Jalankan algoritma *K-Means* dan keluarkan semua *cluster*.

Hasil dari algoritma 2 adalah hasil *clustering* yang jumlah *cluster*nya sudah optimal

2.4 K-Means Clustering

Algoritma *K-Means* digunakan karena *K-Means* merupakan salah satu teknik *clustering* yang paling mudah, cepat, tidak memakan banyak memori, serta bisa diaplikasikan kepada *dataset* berukuran kecil atau besar [14]. Konsep dari algoritma *K-Means* adalah membagi *data points* (*records*) kedalam K jumlah kelompok (*cluster*).

Langkah-langkah dalam algoritma *K-Means* adalah:

1. Tentukan nilai K yang diinginkan. Buat *cluster* sebanyak K.
2. Pilih random *c* dari *dataset* sebagai nilai *centroid* di setiap *cluster* yang ada.
3. Hitung jarak antara *centroid* dengan tiap *record* yang ada di dataset. Untuk menghitung jarak antara *c* dengan tiap record menggunakan rumus *Euclidean Distance*, yaitu [5] :

$$\ldots = \sqrt{\sum_{\ldots=1}^{\ldots} (\ldots(\ldots - \ldots(\ldots)^2}$$

   Dimana $\ldots$ adalah matriks dari objek *i* pada *cluster k,* $\ldots$ adalah koordinat dari objek *j* pada dimensi *k, m* adalah banyak *cluster,* dan $\ldots$ adalah jarak antara objek *i* dan *j*.
4. Masukkan *record* ke dalam *cluster* yang jaraknya paling dekat atau nilai $\ldots$-nya paling kecil.
5. Hitung nilai *centroid* yang baru untuk setiap *cluster* dengan rumus [3] :

$$\ldots = \left(\frac{1}{\ldots}\right) \sum_{\ldots=1}^{\ldots} \ldots$$

   Dimana $\ldots$ adalah nilai dari *record* di dalam *cluster*, $\ldots$ adalah banyak *record* di *cluster* tersebut, dan *c* adalah nilai *centroid* yang baru.

6. Ulangi langkah 3 sampai 5, hingga tidak ada perubahan nilai *centroid*.

3. Perancangan Sistem

Pada bab 3, akan dibahas masing-masing proses dari setiap langkah yang diperlukan untuk menyelesaikan penelitian.

3.1 Pembersihan Data

Dari 162500 baris data tadi terdapat banyak sekali *record* yang tidak dibutuhkan. Oleh karena itu *record* tersebut dihapus. *Record* yang dihapus tersebut adalah :

1. Request yang tidak sukses, yaitu *record* yang memiliki nilai status tidak sama dengan 200
2. Kolom "Identitas Client", "UserID", dan "Ukuran".
3. Request yang mempunyai kata POST, HEAD, /images, /assets/, /plugins/, /favicon.ico, /robots.txt, php, /?utm, /wp. /how, admin, elements, themes, ckeditor, lang=en, activation, dan register.

3.4 Identifikasi Pengguna

Identifikasi Pengguna adalah proses memberikan ID (nomor identitas) yang unik kepada tiap pengguna. Jika IP Address dan agen yang digunakannya sama, maka dihitung sebagai satu pengguna. Dari proses ini ditemukan 4582 UserID yang berbeda.

3.5 Identifikasi Session

Identifikasi Pengguna adalah proses memberikan ID (nomor identitas) yang unik kepada tiap pengguna. Jika IP Address dan agen yang digunakannya sama, maka dihitung sebagai satu pengguna.

3.6 Identifikasi Request

Identifikasi request adalah tahap memberikan nomor unik kepada setiap *request* yang berbeda. Dengan memberi request nomor unik, identifikasi request menemukan 3816 request yang berbeda.

3.7 Identifikasi Kategori

Identifikasi kategori adalah tahap untuk memberikan nomor unik untuk setiap kateogri yang berbeda. Satu kategori bisa memiliki berbagai macam halaman. Dari proses ini ditemukan 11 kategori berbeda, yaitu :

Request	Kategori
GET / HTTP/1.1	Homepage
GET /lariii HTTP/1.1	Proyek
GET /2015/02/inflasi-wacana-vs-crowdfunding/ HTTP/1.1	Blog
GET /about-us HTTP/1.1	Tentang Kitabisa
GET /artikel/17/apa-itu-fitur-penggalang-dana HTTP/1.1	Artikel
GET /dashboard/proyek HTTP/1.1	Dahsboard
GET /dashboard/proyek/buat-baru HTTP/1.1	Buat Proyek
GET /donasi/12191/a4e36c807592bbf/pilih-hadiah HTTP/1.1	Donasi
GET /term-and-conditions HTTP/1.1	Guidance
GET /liputan-media?page=2 HTTP/1.1	Liputan Media

GET /orang-baik/812 HTTP/1.1	Donatur

3.8 Pattern Discovery

Di dalam pattern discovery, dilakukan proses berikut :

1. Sequence Based Analysis
2. *Improved K-Means Clustering*

4. Analisa

Bab ini akan menjelaskan analisa yang dilakukan terhadap hasil *clustering* serta saran perbaikan website.

4.1 Analisa Improved K-Means

Variabel yang menentukan nilai K terbaik dari algoritma *Improved K-Means* adalah α dan βmin. Variabel α ditentukan sendiri oleh pengguna algoritma ini, sedangkan nilai βmin diambil dari rata-rata nilai pada matriks *beta*. Oleh karena itu, dibawah ini akan dilakukan uji skenario algoritma *Improved K-Means* dengan nilai α yang diubah-ubah.

Uji coba akan dilakukan 3 kali. Hasil uji coba adalah sebagai berikut :

α	K dari algoritma 1	βmin	K dari algoritma 2
2.5	12	0.269	6
2.6	11	0.272	6
2.8	5	0	5

Dari 3 buah percobaan terhadap penentuan nilai α, dihasilkan nilai K=6 muncul paling sering sebagai hasil akhir dari algoritma *Improved K-Means.* Oleh karena itu, untuk proses selanjutnya akan digunakan acuan hasil *clustering K-Means* saat K=6.

Berikut adalah hasil *clustering* saat K=6 dan pada saat α = 2.5 :

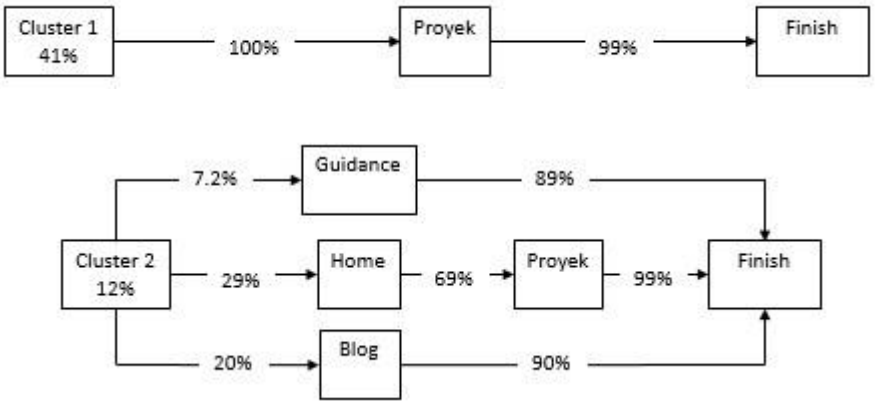
K=6

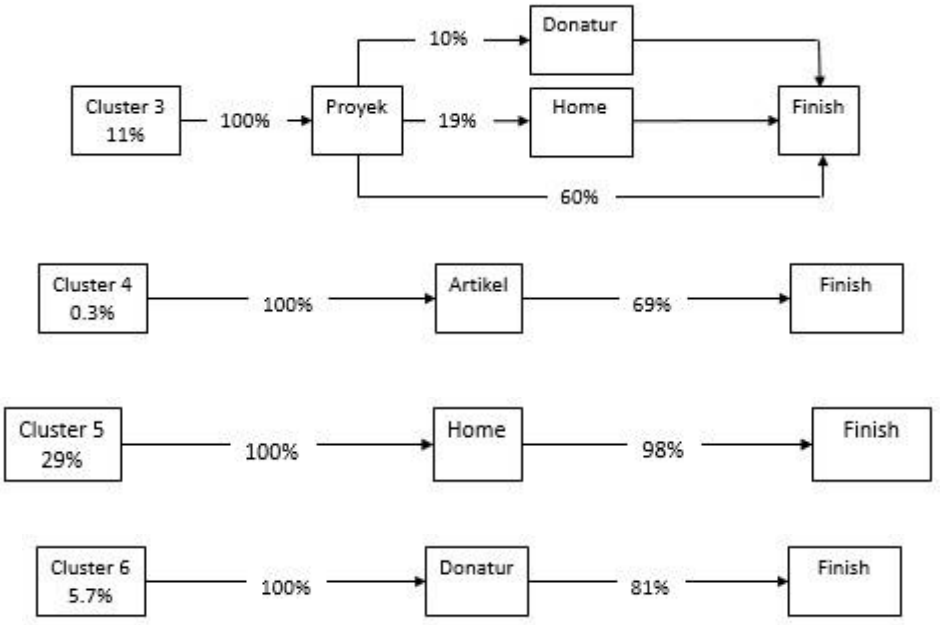
C1	C2	C3	C4	C5	C6
4084	1234	1130	35	2849	567

4.2 Pattern Analysis saat K = 6

Cara kedua untuk mendapatkan hasil *clustering* terbaik adalah dengan menggunakan algoritma *Improved K-means*. Pada tahap sebelumnya, diketahui bahwa jumlah *cluster* terbaik adalah 6, menurut algoritma ini.

Pola pengguna jika jumlah *cluster* adalah 6 :





Dari *cluster* 1 hingga *cluster* 6 diberi identifikasi, C1, C2, C3, C4, C5, dan C6. Detail dari tiap *cluster* ditulis pada tabel berikut :

Cluster ID	Tipe	Jumlah Anggota	Persentase	Rata-rata Aktifitas
C1	Pengguna Proyek	4084	41%	1 cat/sess
C2	Pengguna selain Proyek	1234	12%	1.65 cat/ses
C3	Pengguna selain Home	1130	11%	1.65 cat/ses
C4	Pengguna Artikel	35	0.3%	1.48 cat/ses
C5	Pengguna Home	2849	29%	1 cat/ses
C6	Pengguna Donatur	567	5.7%	1.20 cat/ses

Cluster yang mempunyai anggota paling banyak adalah *cluster* C1, yaitu 41% dari seluruh matriks transisi. Kemudian terbesar kedua adalah C5, yang anggotanya berjumlah 29% dari seluruh matriks transisi. Aktifitas yang paling sering muncul bersamaan adalah mengunjungi halaman home – halaman proyek. Dari C1 hingga C6, diketahui bahwa halaman proyek dan halaman home adalah jenis halaman yang paling banyak dikunjungi. Dan dar C1 hingga C6, halaman yang paling sedikit dikunjungi adalah halaman Liputan Media.

4.3 Kelebihan dan Kekurangan Algoritma Improved K-Means

Setelah mengimplementasikan algoritma *Improved K-Means* kepada *server log* dari Kitabisa.com, maka dapat diketaui kelebihan dan kekurangan algoritma ini, sesuai dengan yang dialami.

Kelebihan :

1. Algoritma *Improved K-Means* langsung membaca data matriks hasil *Sequenced Based Analysis*.

2. *Initial point* yang digunakan untuk *clustering* tidak dipilih secara acak, sehingga hasilnya lebih akurat.

Kekurangan :

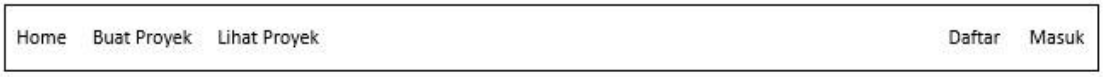
1. Sangat banyak deretan langkah yang ada di dalam algoritma 1 dan algoritma 2.
2. Algoritma 1 dan algoritma 2 cukup kompleks sehingga cukup sulit untuk diubah kedalam bahasa pemrograman.
3. Hasil sangat bergantung kepada variabel α dan βmin, yang keduanya ditentukan dengan keinginan sendiri.

Namun, kekurangan algoritma *Improved K-Means* dimana hasil sangat bergantung kepada dua variabel yang dipilih secara acak, bisa diperbaiki dengan cara mendapatkan nilai α dari jarak rata-rata antar matriks, dan nilai βmin didapat dari rata-rata faktor β(i,j). Terbukti hasil yang diberikan *Imporved K-Means* dengan perbaikan tersebut, memberikan hasil yang bagus.

4.4 Saran Perbaikan Website

Dilihat dari pola akses pengguna di C1, C2, C3, dan C6, terlihat bahwa jumlah kunjungan ke halaman liputan media adalah yang paling sedikit. Hal ini bisa disebabkan oleh beberapa faktor, misalkan : halaman tidak menarik, link menuju halaman tersebut tidak ada, atau memang halaman tersebut tidak diperlukan di dalam sistem ini. Jika pengembang website ingin memperbaiki sistem agar kunjungan ke halaman liputan media, maka berikut ini adalah saran perbaikan kerangka tampilan websitenya.

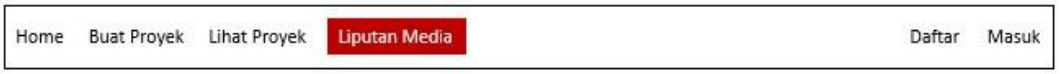
Rangka tampilan Bar Navigasi (Navbar) saat ini :



Gambar Error! No text of specified style in document.**.1 Kerangka tampilan navigasi bar pada Kitabisa.com**

Menambah menu navigasi Liputan Media, namun yang lebih menarik daripada menu navigasi yang lain. Misalnya dengan memberi latar belakang warna yang berbeda.

Rangka tampilan Navbar yang disarankan :



Gambar 4.1 Saran perbaikan navigasi bar

5. Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan beberapa hal berikut :

1. Aspek yang bisa diperbaiki dari melihat hasil pola akses pengguna adalah memperbaiki sedikitnya kunjungan ke halaman liputan media.
2. Hasil dari algoritma *Improved K-Means* untuk kasus Kitabisa.com adalah nilai K terbaik sama dengan 6.
3. Perbaikan saran yang diberikan untuk menyocokkan aktifitas pengguna dengan tampilan. Halaman yang diperbaiki oleh saran adalah bagian menu navigasi.

Daftar Pustaka

[1]SungJune, P. (2008). Sequence-based *cluster*ing for Web usage mining: A new experimental framework and ANN-enhanced. *Elsevier*.

[2] Agrawal, M., & Mishra, M. (2009). Improved K-Mean *Cluster*ing Approach for Web Usage Mining. *Second International Conference on Emerging Trends in Engineering and Technology*.

[3] Alsabti, K., Ranka, S., & Singh, V. (n.d.). An Eddicient K-Means *Cluster*ing Algorithm.

[4] Hedresta, T., Darwiyanto, E., & Effendy, V. (2014). Analisi dan Implementasi Web Usage Mining Menggunakan Metode Self Organizing Map dan K-Means (Studi Kasus : Aktifitas Internet Telkom University).
[5] Kopengguna, K., & Sunita. (2013). A comparative study of K Means Algorithm. *International Journal of Innovative Research in Computer*.
[6] Vijayashro, L . & Joshi, M. (2012). Data Preprocessing in Web Usage Mining. *International Conference on Artificial Intelligence and Embedded Systems* .
[7] Hamerly, Greg. & Charles, Elkan. (2004). Learning The K in K-Means.
[8] Srivasta, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data.
[9] Verma, M., Srivastava, M., Chack, N., Diswar, K. A., & Gupta, N. (2012). A Comparative Study of Various *Cluster*ing Algorithms in Data Mining. 1779-1384.
[10] culture-ist. (2013, May 9). *More Than 2 Billion People Use the Internet, Here's What They're Up To*. Retrieved from The CultureIst: http://www.thecultureist.com/2013/05/09/how-many-people-use-the-internet-more-than-2-billion-infographic/
[11] Kickstarter. http://www.kickstarter.com
[12] Yu-Shiang Hung, Kuei-Ling B.Chen, Chi-Ta Yang, Guang-Feng Deng. (2013). Web Usage Mining for Analysing Elfer Self-care Behavior Patterns. *Expert Systems with Applications 40 (2013) 775–783*
[13] Datasciencelab. (2014, Januari 21). *Selection of K in K-Means Clustering, Reloaded.* The Data Science Lab : https://datasciencelab.wordpress.com/2014/01/21/selection-of-k-in-k-means-*cluster*ing-reloaded/
[14] Zeferino. (2013, Mei 13). *Why do we use k-means instead of other algorithms?*. Stack Exchange : http://stats.stackexchange.com/questions/58855/why-do-we-use-k-means-instead-of-other-algorithms/
[15] Wikipedia. *Web Browser*. Wikipedia : https://en.wikipedia.org/wiki/Web_browser. Akses : 12 Januari 2016
[16] Wikipedia. *Alamat IP*. Wikipedia : https://id.wikipedia.org/wiki/Alamat_IP. Akses : 12 Januari 2016
[17] Wikipedia. *K-Means Clustering*. Wikipedia : https://en.wikipedia.org/wiki/K-means_clustering. Akses : 12 Januari 2016
[18] Wikipedia. *Markov Chain* . Wikpedia : https://en.wikipedia.org/wiki/Markov_chain. Akses : 12 Januari 2016
[19] Wikipedia. *Server (computing).* Wikipedia : https://en.wikipedia.org/wiki/Server_(computing). Akses : 12 Januari 2016
[20] Wikipedia. *Website*. Wikipedia : https://en.wikipedia.org/wiki/Website. Akses : 12 Januari 2016
[21] Wikipedia. *XML*. Wikipedia : https://en.wikipedia.org/wiki/XML. Akses : 12 Januari 2016
[22] Dhillon, Supreet. (2014). Comparative Study of Classification Algorithms for Web Usage Mining.