PENGELOMPOKAN PENGGUNA SITUS WEB BPS MELALUI TEKNIK *BIBLIOMETRIC* DAN ANALISIS KORESPONDENSI

Toza Sathia Utiayarsih[1], Jadi Suprijadi[2], Bernik Maskun[3]

[1]Politeknik Statistika STIS
[2,3]Universitas Padjajaran
e-mail: [1]toza@stis.ac.id

Abstrak

Salah satu upaya pemenuhan program percepatan (*quick wins*) terhadap produk BPS yang benar-benar dapat menyentuh kebutuhan para pengguna data adalah dengan melakukan segmentasi terhadap pengguna data. Segmentasi terhadap pengguna situs web BPS sebagai salah satu bentuk segmentasi terhadap pengguna data, sesuai program percepatan. Ukuran data pengguna web sangat besar dan berupa data teks sehingga tidak dapat langsung dianalisis melalui aplikasi statistik yang tersedia, maka perlu dilakukan suatu teknik untuk data pengguna web dengan menggunakan teknik *bibliometric*. Teknik tersebut mengubah data teks menjadi format numerik, selanjutnya dibuat menjadi matriks distribusi frekuensi. Matriks digunakan pada analisis korespondensi untuk pengelompokkan pengguna situs web. Hasil dari analisis pengguna situs web BPS yang diwakili oleh alamat IP dapat dikelompokkan dengan halaman yang diakses berdasarkan asal negara, sehingga didapatkan segmentasi pengguna data situs web BPS antara negara dan halaman yang diakses.

Kata kunci: *Data Mining, text mining, bibliometric, web mining*, analisis korespondensi

Abstract

The effort to fulfill one of quick wins program for BPS products that really can fulfill the needs of data users is by segmenting data users. Segmentation of BPS website users as a form of segmentation of data users, according to quick wins program. The size of web user data is very large and in the form of text data so that it cannot be directly analyzed through available statistical applications, it is necessary to do a technique for web user data using bibliometric techniques. This technique converts text data into numeric format, then it is made into a frequency distribution matrix. The matrix is used in correspondence analysis for grouping website users. The results of the analysis of BPS website users represented by IP addresses can be grouped with pages accessed based on national origin, so that segmentation users of BPS website data between the country and the page are accessed can be obtained.

***Keywords**: Data Mining, text mining, bibliometric, web mining, correspondence analysis*

PENDAHULUAN

Badan Pusat Statistik (BPS) selalu berupaya untuk melakukan perubahan dan reformasi yang mendasar terhadap sistem penyelenggaraan kegiatan statistik, melalui pembangunan profil dan perilaku aparatur BPS yang profesional, berintegritas, bertanggung jawab, serta mampu memberikan pelayanan prima kepada publik. BPS sebagai lembaga pemerintah non-kementerian mempunyai tugas untuk menyediakan data dan informasi statistik yang berkualitas, serta dituntut untuk melayani berbagai kepentingan pengguna data. Sejalan dengan keinginan reformasi birokrasi, ke depan BPS harus mampu menghasilkan data yang berkualitas, yang didukung oleh SDM profesional dan infrastruktur yang lebih modern.

Untuk membangun kepercayaan masyarakat perlu diupayakan suatu program percepatan (*quick wins*) terhadap produk BPS yang benar-benar dapat menyentuh kebutuhan para pengguna data. Program *quick wins* ini dipilih dengan memperhatikan produk statistik yang memiliki daya ungkit tinggi, inovatif, dan merupakan terobosan yang terkait dengan produk utama BPS. Program *quick wins* yang memenuhi kriteria tersebut di atas antara lain: (i). Peningkatan Kepuasan Pelanggan, (ii). Penyempurnaan Pelayanan Statistik yang terdiri pelayanan Elektronik (*e-Services*) dan pelayanan statistik terpadu yang menggabungkan pelayanan perpustakaan (digital dan non-digital), konsultasi statistik, toko buku (*e-Shop*) dan pelayanan lainnya, dan (iii). Membangunan *Advanced Release Calendar*.

Dalam upaya memenuhi kriteria tersebut muncul salah satu tujuannya yaitu segmentasi pengguna data baik melalui pelayanan langsung maupun pelayanan elektronik (*e-Service*) seperti situs web BPS (Laporan Reformasi Birokrasi Badan Pusat Statistik, 2011). Sejalan dengan hal tersebut, perlu diketahui tentang pola pengguna situs web itu sendiri dalam rangka mendapatkan segmentasi pengguna yang tepat. Untuk menganalisis pola pengguna situs web dibutuhkan suatu instrumen yang dapat menjembatani antara pengguna dengan pengelola situs web, yaitu melalui *web usage session*, yang merupakan interaksi antara pengguna dan *web server* dalam satu periode waktu tertentu yang berisi halaman web yang dikunjungi.

Data mining merupakan pendekatan yang sangat berguna pada aspek pengolahan data dan penelaahan penemuan. Pada dasarnya, data mining mengacu pada ekstraksi informasi data dalam jumlah besar, yang memiliki berbagai macam bentuk atau jenis data, seperti data transaksi pada aplikasi web (pembelian online, layanan konsumen, dll). Dalam sepuluh tahun terakhir, menurut Xu (2010), *data mining* berhasil masuk ke dalam dunia penelitian manajemen data web, seperti dokumen web, struktur tautan web, transaksi pengguna web, dan web semantics menjadi target penelaahan. Jelas bahwa informasi yang dapat digali dari berbagai jenis data web dapat membantu dalam menemukan hubungan antara berbagai obyek dalam web sehingga dapat meningkatkan manajemen data web.

Menurut wikipedia, *web mining* merupakan suatu aplikasi bagian dari *data mining* yang menggali pola-pola yang tersedia di dalam web itu sendiri. Jadi antara *data mining* dan *web mining* hanya berbeda dalam hal target data yang dianalisis. Data mining umumnya menganalisis data yang berasal dari OLTP (Online Transactional Process) dan data transaksi lainnya. Sedangkan *web mining* target analisisnya adalah data dari web, seperti data akses pengunjung, struktur halaman web, format halaman web dan sebagainya. Berdasarkan target analisisnya, *web mining* dibagi menjadi 3 (tiga) bagian, yaitu: (i). *web content mining*, (ii). *web structure mining*, dan (iii). *web usage mining*.

Menurut Srivastava (2000), *web usage mining* merupakan teknik *data mining* yang menggambarkan pola penggunaan dari halaman web, dalam rangka memahami dan meningkatkan pelayanan kebutuhan dari aplikasi berbasis web. Sumber data utama dari *web usage mining* adalah *server logs* dan *browser logs*.

Tabel 1. Kategori Halaman yang Diberi Label Kode Angka

Nama Halaman Web	Kode	Nama Halaman Web	Kode
Beranda	1	Publikasi BPS	9
Tentang BPS	2	Berita Resmi BPS	10
Rencana Strategis BPS	3	Unduh	11
Pusat Layanan	4	Berita	12
Istilah Statistik	5	Info Lelang	13
Jabatan Fungsional	6	Subyek Statistik	14
Sistem Rujukan Statistik	7	*Website* BPS Provinsi	15
Sekolah Tinggi Ilmu Statistik	8		

Teknik *server log analysis* digunakan jika memiliki akses penuh terhadap suatu situs web dan *server* web yang digunakan. Karena data tersimpan di dalam file, maka data log relatif mudah dikelola. Data yang tercatat pada *log server* memiliki format teks dalam jumlah yamg sangat besar. Data tersebut merupakan data tidak terstruktur, tetapi memungkinkan untuk diubah menjadi bentuk bibliografi sehingga bisa diterapkan metode untuk mengolahnya melalui teknik *bibliometric*. *Software bibliometric* sebagai alat untuk analisis informasi dalam jumlah yang besar berkembang dengan *output* format yang bervariasi, misalnya, distribusi frekuensi, matriks, peta, dan network (Supriyadi, 2011).

Ukuran data web sangat besar dan berupa data teks sehingga tidak dapat langsung dianalisis melalui *software* statistik biasa, maka perlu dilakukan suatu teknik untuk data pengguna web sehingga dapat berubah menjadi format numerik, seperti matriks distribusi frekuensi yaitu melalui teknik *bibliometric*. Selanjutnya hasil yang didapatkan melalui *bibliometric* dapat digunakan pada analisis statistik untuk pengelompokkan pengguna situs web BPS sehingga dapat dilihat pola segmentasi dari pengguna.

METODE

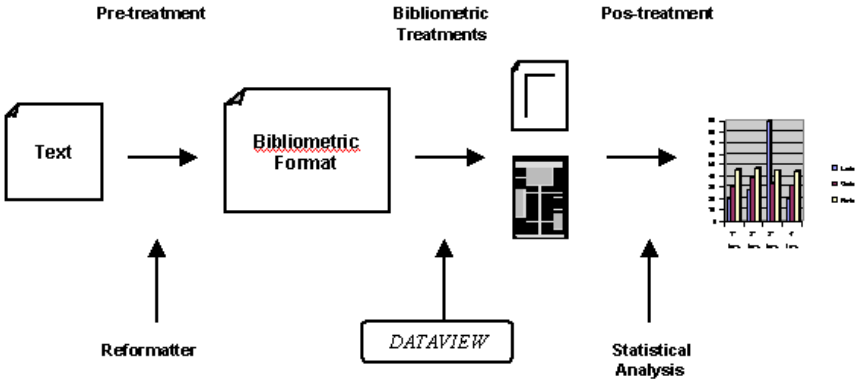
Berdasarkan uraian permasalahan yang disampaikan pada pendahuluan, dapat dirumuskan dalam penelitian ini adalah bagaimana mengolah data pengguna web pada halaman situs web BPS dengan format teks menjadi numerik, sehingga dapat dilakukan analisis terhadap data tersebut, dan diketahui pola segmentasi pengguna melalui salah satu analisis statistik. Dengan tahapan mengubah data pengguna situs web BPS yang berbentuk teks menjadi format numerik melalui teknik *bibliometric* yang dapat menghasilkan matriks kontingensi. Kemudian matriks tersebut bisa dilanjutkan dengan analisis statistik menggunakan analisis korespondensi. Sehingga didapatkan pola segmentasi pengguna situs web BPS melalui pengelompokkan berdasarkan asal negara.

1. Data *Web Usage* Situs Web BPS

Kategori halaman yang digunakan pada penelitian ini berasal dari peta situs web BPS yang merupakan kerangka dasar dalam sebuah situs web yang berisi informasi mengenai halaman-halaman yang ada dalam situs. Halaman pada situs web BPS terdapat 15 kategori, yaitu: “Beranda”, “Tentang BPS”, “Rencana Strategis BPS”, “Pusat Layanan”, “Istilah Statistik”, “Jabatan Fungsional”, “Sistem Rujukan Statistik”, “Sekolah Tinggi Ilmu Statistik”, “Publikasi BPS”, “Bertita Resmi BPS”, “Unduh”, “Berita”, “Info Lelang”, “Subyek Statistik”, “Website BPS Provinsi”.

Setiap kategori halaman direpresentasikan dengan label *integer*. Contohnya, “Beranda” diberi kode 1, “Tentang BPS” diberi kode 2, “Rencana Strategis BPS” diberi kode 3, dan seterusnya, seperti terlihat dalam Tabel 1.

Sumber data sebagian besar *web usage mining* adalah *web server log*, yang menyediakan data mentah untuk



Gambar 1. Posisi Dataview dalam Rantai Pengolahan *Bibliometric*

(Sumber: Rostaing, 2000, dalam Tarapanoff et al, 2001)

mengidentifikasi kumpulan data web atau web usage session. *Web server log* berisi catatan akses dari pengguna. Setiap *record* mewakili sebuah halaman yang diakses oleh pengguna dan umumnya berisi alamat IP (*Internet Protocol*) pengguna, tanggal dan waktu akses diterima, alamat URL yang diakses, kode balasan dari *server* yang menunjukkan status akses, dan ukuran file (byte) dari halaman yang diakses sempurna.

2. Teknik *Bibliometric*

Data text yang didapatkan dianalisis dengan menggunakan proses *bibliometric*. Tahapan Teknik *bibliometric* seperti yang dapat dilihat pada Gambar 1 adalah sebagai berikut:

1. Data web berupa text file yang tidak terstruktur diubah menjadi database terstruktur. Data yang diambil dari web log pada server perlu disiapkan sebelum memasuki proses pengolahan atau biasa disebut sebagai *preprocessing*. Proses ini terdiri dari 2 (dua) tahapan, yaitu: pemilihan data dan transformasi data menjadi data yang terstruktur. Hasil dari proses ini adalah database web server log.
2. *Database web server log* terdiri dari 5 field, yaitu: *Internet Protocol* (IP), waktu, halaman, status, dan ukuran.
3. Data Internet Protocol (IP) dan halaman ditransformasi menjadi format *bibliometric*. Data format *bibliometric* terdiri dari field nomor record (NO), alamat IP pengguna/*Internet Protocol* (IP), dan halaman yang diakses (HAL). Data ini berupa text file sehingga lebih mudah dikelola dalam proses bibilometric.
4. Data format *bibliometric* diolah dengan proses bibilometric sehingga menghasilkan *output* tabel kontingensi dengan baris adalah field Internet Protocol (IP) dan kolom adalah field halaman.
5. Tabel kontingensi disederhanakan dengan mengklasifikasikan alamat IP pengguna/ *Internet Protocol* (IP) berdasarkan negara.
6. Tabel kontingensi yang telah disederhanakan kemudian dianalisis dengan menggunakan analisis statistik.

3. Analisis Korespondensi

Tabel kontingensi yang dihasilkan melalui teknik *bibliometric* kemudian dianalisis dengan menggunakan analisis statistik, dalam penelitian ini digunakan analisis korespondensi sederhana. Menurut Izenman (2008), proses dari analisis sebagai berikut:

Tabel Kontingensi Dua Arah

Data kategorik adalah data yang dikumpulkan dari hasil hitungan yang disusun dalam tabel kontingensi. Sebuah tabel kontingensi dua arah ($r \times s$) dengan r baris (diberi label $A_1, A_2, \ldots, A_r$) dan s kolom (diberi label $B_1, B_2, \ldots, B_s$) terdiri dari rs sel. Sel ke-ij, n_{ij}, mewakili frekuensi yang diamati untuk baris kategori A_i dan kolom kategori B_j, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, s$. Total marjinal baris ke-i adalah $n_{i+} = \sum_{j=1}^{s} n_{ij}$, $i = 1, 2, \ldots, r$, dan total marjinal

Tabel 2. Tabel Kontingensi Dua Arah yang Menjelaskan Frekuensi Sel Pengamatan, Total Marjinal Baris & Kolom, dan Jumlah Sampel

Variabel Baris	Variabel Kolom						Total Baris
	B_1	B_2	…	B_j	…	B_s	
A_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	n_{1+}
A_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	n_{2+}
…	…	…		…		…	…
A_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	n_{i+}
…	…	…		…		…	…
A_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	n_{r+}
Total Kolom	n_{+1}	n_{+2}	…	n_{+j}	…	n_{+s}	n_{++}

kolom ke-j adalah $n_{+j} = \sum_{i=1}^{r} n_{ij}$, *j = 1, 2, ... , s*. Jika $n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$ individu diklasifikasikan oleh kategori baris dan kolom, kemudian Tabel 3, yang juga disebut tabel korespondensi, menunjukkan frekuensi sel, total marjinal, dan total ukuran sampel.

Notasi π_{ij} merupakan peluang bahwa seorang individu memiliki karakteristik A_i dan B_j, *i = 1, 2, ... , r, j = 1, 2, ... , s*. Dengan asumsi bahwa baris variabel *A* dan kolom variabel *B* adalah independen, sehingga $\pi_{ij} = \pi_{i+}\pi_{+j}$, dengan $\pi_{i+} = \sum_j \pi_{ij}$ dan $\pi_{+j} = \sum_i \pi_{ij}$, untuk semua *i = 1 2, ... , r* dan *j = 1, 2, ... , s*. Secara umum yang ingin dilihat adalah apakah *A* dan *B* memang variabel independen. Sebuah pertanyaan dapat diajukan sebagai alternatif dalam hal homogenitas dari distribusi peluang baris atau kolom, yaitu, apakah semua baris memiliki distribusi peluang yang sama di setiap kolom, atau sebaliknya, semua kolom memiliki distribusi peluang yang sama di setiap baris.

Variabel Dummy Baris dan Kolom

Pada tabel kontingensi dua arah, dapat melihat hubungan antara kategori baris dan kategori kolom seperti pada Tabel 2.

Merubah tabel kontingensi N menjadi "matriks korespondensi" sebagaimana Tabel 3

Jarak Chi-Square

Pada analisis korespondensi, penting untuk menggambarkan jarak diantara profil baris (yaitu baris pada matriks Pr) atau diantara profil kolom (yaitu kolom pada matriks Pc). Untuk mengukur jarak ini digunakan ukuran chi-squared.

1. Jarak Baris

Jika profil baris ke-i dan ke-i' adalah **ai** dan **ai'**, maka **ai - ai'** adalah s-vektor dengan elemen ke-j $n_{ij}/n_{i+} - n_{i'j}/n_{i'+}$. Kuadrat dari jarak *chi-squared* diantara $\mathbf{a_i}$ dan $\mathbf{a_i'}$ sebagai berikut:

$$d^2(a_i, a_i') = (a_i - a_i')^T D_c^{-1}(a_i - a_i')$$

$$= \sum_{j=1}^{s} \frac{n}{n_{+j}}\left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i'+}}\right)^2 \quad (1)$$

Perhatikan Persamaan (1), massa kolom ke-j (n_{+j}/n) masuk ke dalam persamaan tersebut berbanding terbalik dengan kuadrat jarak dari profil baris. Sehingga jumlah observasi (*n*) berpengaruh terhadap jarak antar profil baris.

Perhatikan bahwa **c** adalah sentroid baris. Matriks berukuran (*r x s*) dari titik pusat profil baris $\mathbf{P_r} - \mathbf{1_r c^T}$ dengan $\mathbf{P_r} = \mathbf{D_r^{-1}P}$, memiliki baris ke-I $(\mathbf{a_i} - \mathbf{c})^\mathbf{T}$ dengan elemen ke-j $n_{i+}^{-1}(n_{ij} - n_{i+}n_{+j}/n)$, *i = 1, 2, ... , r, j = 1, 2, ... , s*. Sehingga kuadrat dari jarak *chi-squared* antara $\mathbf{a_i}$ dan **c** adalah:

$$d^2(a_i, c) = (a_i - c)^T D_c^{-1}(a_i - c)$$

$$= \frac{1}{n_{i+}}\sum_{j=1}^{s} \frac{n}{n_{i+}n_{+j}}\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2$$

(2)

Tabel 3. Matriks Korespondensi Menjelaskan Frekuensi Relatif dari Sel Pengamatan, Total Marjinal Baris, dan Total Marjinal Kolom terhadap n

Variabel Baris	**Variabel Kolom**						**Total Baris**
	B_1	B_2	…	B_j	…	B_s	
A_1	p_{11}	p_{12}	…	p_{1j}	…	p_{1s}	p_{1+}
A_2	p_{21}	p_{22}	…	p_{2j}	…	p_{2s}	p_{2+}
…	…	…		…		…	…
A_i	p_{i1}	p_{i2}	…	p_{ij}	…	p_{is}	p_{i+}
…	…	…		…		…	…
A_r	p_{r1}	p_{r2}	…	p_{rj}	…	p_{rs}	p_{r+}
Total Kolom	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+s}	1

Penjumlahan dari semua profil baris pada Persamaan (2) menjadi:

$$n\sum_{i=1}^{r} p_{i+}d^2(a_i,c) = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(n_{ij}-\frac{n_{i+}n_{+j}}{n}\right)^2 \Big/ \left(\frac{n_{i+}n_{+j}}{n}\right) \quad (3)$$

dengan statistik uji *Pearson Chi-Squared* sebagai berikut:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (4)$$

Frekuensi sel hasil observasi O_{ij} dan frekuensi sel harapan E_{ij} (dengan asumsi baris dan kolom independen) sebagai berikut:

$$O_{ij} = n_{ij}, \; E_{ij} = \frac{n_{i+}n_{+j}}{n} \quad (5)$$

dengan *i = 1, 2, ... , r, j = 1, 2, ... , s*. Di bawah asumsi sampel acak, χ^2 pada Persamaan (4) mendekati *distribusi chi-squared* (χ^2) pada sampel besar dengan derajat bebas (*r*-1)(*s*-1).

2. Jarak Kolom

Sama seperti jarak baris, untuk jarak kolom, jika profil kolom ke-j dan ke-j' adalah $\mathbf{b_j}$ dan $\mathbf{b_j'}$, maka $\mathbf{b_j} - \mathbf{b_j'}$ adalah *r*-vektor dengan elemen ke-j $n_{ij}/n_{+j} - n_{ij'}/n_{+j'}$. Kuadrat dari jarak *chi-squared* diantara $\mathbf{b_j}$ dan $\mathbf{b_j'}$ sebagai berikut:

$$d^2(b_j,b_j') = (b_j-b_j')^T D_r^{-1}(b_j-b_j')$$

$$= \sum_{i=1}^{r} \frac{n}{n_{i+}}\left(\frac{n_{ij}}{n_{+j}}-\frac{n_{ij'}}{n_{+j'}}\right)^2 \quad (6)$$

Kuadrat dari jarak *chi-squared* antara $\mathbf{b_j}$ dan **r** adalah:

$$d^2(b_j,r) = (b_j-r)^T D_r^{-1}(b_j-r)$$

$$= \frac{1}{n_{+j}}\sum_{i=1}^{r} \frac{n}{n_{i+}n_{+j}}\left(n_{ij}-\frac{n_{i+}n_{+j}}{n}\right)^2 \quad (7)$$

Penjumlahan dari semua profil kolom pada Persamaan (7) menjadi:

$$n\sum_{j=1}^{s} p_{+j}d^2(b_j,r) = \chi^2 \quad (8)$$

dengan χ^2 seperti pada Persamaan (4).

Sehingga rata-rata tertimbang dari kuadrat jarak *chi-squared* pada semua profil baris terhadap sentroid baris atau pada semua profil kolom terhadap sentroid kolom (dengan penimbang massa baris/massa kolom) adalah χ^2/n. Jika baris dan kolom independen, maka χ^2/n akan kecil, sejalan dengan $p_{i+}d^2(a_i,c)$ dan $p_{+j}d^2(b_j,r)$.

Di sisi lain, jika χ^2/n besar, berarti minimal ada satu dari $p_{i+}d^2(a_i,c)$ atau $p_{+j}d^2(b_j,r)$ akan besar. Informasi ini penting untuk menentukan apakah independensi dalam tabel terpenuhi atau tidak. Bandingkan matriks tersebut dengan matriks **N** = (O_{ij}).

Total Inersia

Dengan menggunakan dummy variable untuk mewakili tabel kontingensi dua arah, memungkinkan untuk melihat suatu masalah sebagai suatu kasus khusus dari analisis kanonik. Bagaimanapun situasinya berbeda, bahwa apabila menggali struktur korelasi antara dua set dari vektor data statistik, akan berhadapan dengan struktur korelasi dari dua set dummy variable.

Tabel 4. Struktur Data *Web Server Log* Hasil Pemilihan Data

Nama Field	Deskripsi	Tipe Data
IP	Alamat IP Pengguna	*Text*
WAKTU	Tanggal dan Jam Akses	*Date*
URL	URL yang Diakses	*Text*
STATUS	Status Akses	*Numeric*
UKURAN	Ukuran Halaman yang Diakses	*Numeric*

Jika nilai dari χ^2 sangat besar, asumsi independensi dari variansi baris dan kolom pada tabel kontingensi tidak terpenuhi (ditolak). Selanjutnya menentukan dimana deviasi dari keindependenan terjadi. Nilai dari χ^2/n mengacu pada nilai total inersia pada tabel kontingensi. Nilai inersia utama merupakan persentase dari total variansi yang dijelaskan oleh beberapa komponen utama, yang biasanya terdiri dari 2 (dua) atau 3 (tiga) komponen utama.

Tampilan Grafis

Pada analisis korespondensi, dapat dipilih hanya dengan menganalisis profil baris atau profil kolom, atau menganalisis keduanya. Tampilan grafis dibentuk dengan membuat plot dari koordinat baris dan koordinat kolom yang merupakan *scatterplot*. Tampilan grafis terdiri dari 2 (dua) jenis, yaitu:

1. *Symetric map*: Baik koordinat baris dan koordinat kolom, keduanya dianggap sebagai koordinat utama.
2. *Asymetric map*: Koordinat baris (atau kolom) dianggap sebagai koordinat utama, sedangkan yang lainnya dianggap sebagai koordinat biasa.

Secara garis besar, titik yang terlihat dekat diantara satu sama lain menunjukkan hubungan antar kategori. Lebih jelasnya sebagai berikut:

1. Jika titik pada baris dekat, maka baris tersebut memiliki distibusi bersyarat yang sama pada setiap kolom.
2. Jika titik pada kolom dekat, maka kolom tersebut memiliki distibusi bersyarat yang sama pada setiap baris.
3. Jika titik pada baris dan kolom dekat, maka hal tersebut menyatakan bahwa deviasi tertentu dari independensi atau baris dan kolom menyimpang dari independensi.

HASIL DAN PEMBAHASAN

1. *Preprocessing* Data

Preprocessing data terdiri dari 2 (dua) tahapan, yaitu: pemilihan data dan transformasi data.

Pemilihan Data

Data web log pada server situs web BPS memiliki ukuran yang sangat besar dan berupa text file. Sehingga perlu ditentukan batasan dari segi waktu untuk analisis data pada penelitian ini. Pada penelitian ini, data yang dianalisis adalah data web server log bulan November 2011. Karena keterbatas *software*, data yang diproses adalah data 3 (tiga) hari pada bulan tersebut, yaitu tanggal 1 (satu), 2 (dua) dan 3 (tiga). Data text file kemudian dimasukkan ke dalam *database web server log* agar menjadi file yang terstruktur. Pada tahapan ini dilakukan juga proses *cleaning data* untuk menghilangkan data yang berulang (redundant) dan pemilihan data yang berstatus berhasil melakukan akses. Ukuran *database* untuk 3 hari sebanyak 61.759 *record*. Struktur data *web server log* hasil pemilihan data dapat dilihat pada Tabel 4.

Proses pemilihan data menggunakan program yang dirancang dengan menggunakan bahasa pemrograman Microsoft Visual Basic.NET yaitu melalui fasilitas tombol "Cleaning".

Tranformasi Data

Data *web server log* yang sudah dipilih masih belum sesuai dengan struktur data untuk analisis pada penelitian ini. Struktur data yang dimaksud adalah

Tabel 5. Struktur Data Hasil Transformasi

Nama Field	Deskripsi	Tipe Data
IP	Alamat IP Pengguna	*Text*
HALAMAN	Halaman yang Diakses	*Text*

Tabel 6. Matriks Kontingensi Hasil dari *Bibliometric*

IP	11	14	1	10	2	9	12	5	13	4	3	6	15	8
193.130.130.153	87	2342	916	280	1	0	0	0	0	0	0	0	0	0
223.255.225.75	2398	0	0	0	0	0	0	0	0	0	0	0	0	0
50.115.185.87	1524	108	499	0	0	2	0	0	0	0	0	0	0	0
66.249.69.24	68	422	10	616	174	40	168	17	6	1	10	1	1	1
69.191.249.202	0	548	132	563	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
103.10.169.235	1	0	0	0	0	0	0	0	0	0	0	0	0	0
101.255.16.202	1	0	0	0	0	0	0	0	0	0	0	0	0	0
10.5.3.21	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1.113.17.82	0	0	1	0	0	0	0	0	0	0	0	0	0	0

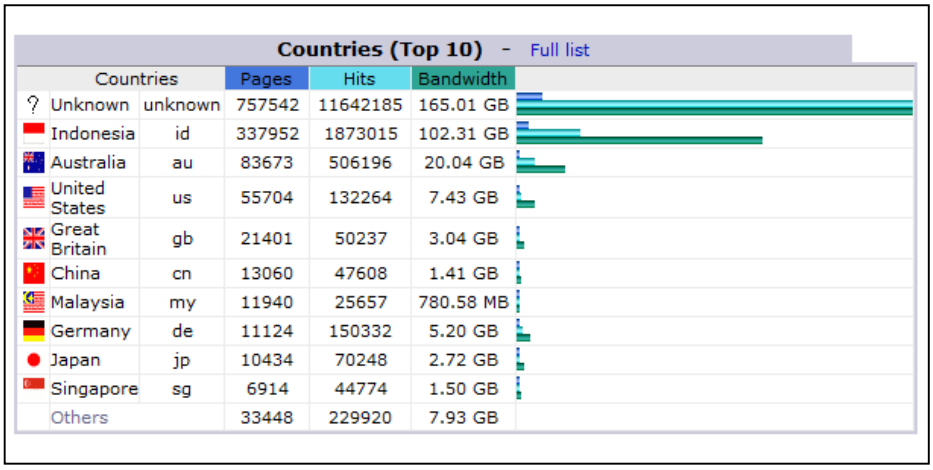
struktur data yang menggambarkan pola akses data berdasarkan halaman yang diakses. Sehingga perlu dilakukan proses transformasi data, yaitu dengan mentransformasi alamat URL yang diakses menjadi halaman-halaman yang ada dalam situs. Halaman pada situs web BPS terdapat 15 kategori, yaitu: "Beranda", "Tentang BPS", "Rencana Strategis BPS", "Pusat Layanan", "Istilah Statistik", "Jabatan Fungsional", "Sistem Rujukan Statistik", "Sekolah Tinggi Ilmu Statistik", "Publikasi BPS", "Berita Resmi BPS", "Unduh", "Berita", "Info Lelang", "Subyek Statistik", "Website BPS Provinsi". Setiap kategori halaman direpresentasikan dengan label integer. Contohnya, "Beranda" diberi kode 1, "Tentang BPS" diberi kode 2, "Rencana Strategis BPS" diberi kode 3, dan seterusnya.

Struktur data hasil transformasi dapat dilihat pada Tabel 5.

Data ditransformasi menjadi bentuk IP berdasarkan halaman web yang diakses. Setelah data terbentuk, diperhatikan bahwa terdapat beberapa pengguna memiliki karakteristik khusus, yaitu pengguna yang mengakses langsung pada halaman tertentu dan mengaksesnya berulangkali. Data web server log hasil transformasi data kemudian dirubah lagi menjadi format bibliografi yang kemudian akan digunakan dalam teknik biblometrik. Data diubah ke dalam field-field yang berisi form dalam format text.

2. Penerapan Teknik *Bibliometric*

Berdasarkan data yang telah dirubah bentuknya menjadi format bibliografi, maka selanjutnya diterapkan teknik *bibliometric* untuk mendapatkan bentuk yang dapat dianalisis lebih lanjut, dari format aslinya yang berupa teks. Pada tahap ini, data diolah dengan *software* khusus untuk format bibliografi. Selanjutnya data dalam dalam format bibliografi diubah bentuknya oleh *software* menjadi field "IP" (alamat IP) dan "HAL" (Halaman yang Diakses), sedangkan isi dari field tersebut menjadi form yang kemudian akan diekstrak dan dipasangkan (pair) antar field melalui suatu proses hingga menghasilkan matriks kontingensi dua arah berukuran 2.618 baris yang merepresentasikan pengguna, dalam hal ini IP, dan 14 kolom yang merepresentasikan halaman yang diakses, dalam hal ini kategori halaman seperti pada Tabel 1. Matriks tersebut dapat dilihat pada Tabel 6.



Countries (Top 10) - Full list

Countries		Pages	Hits	Bandwidth
Unknown	unknown	757542	11642185	165.01 GB
Indonesia	id	337952	1873015	102.31 GB
Australia	au	83673	506196	20.04 GB
United States	us	55704	132264	7.43 GB
Great Britain	gb	21401	50237	3.04 GB
China	cn	13060	47608	1.41 GB
Malaysia	my	11940	25657	780.58 MB
Germany	de	11124	150332	5.20 GB
Japan	jp	10434	70248	2.72 GB
Singapore	sg	6914	44774	1.50 GB
Others		33448	229920	7.93 GB

Gambar 2. 10 Negara Tertinggi yang Mengakses Situs Web BPS

Tabel 7. Kode Negara Klasifikasi Alamat IP

Negara	Kode	Negara	Kode
Indonesia	1	Malaysia	6
Australia	2	Jerman	7
USA	3	Jepang	8
Inggris	4	Singapura	9
Cina	5	Lainnya	10

Kategori halaman yang muncul pada matriks ini hanya 14 dari keseluruhan 15 kategori, hal ini disebabkan salah satu kategori tersebut, yaitu kode "7" tidak ada yang mengakses dalam 3 hari data yang dimasukkan ke dalam pengolahan. Pada *software* apabila isian kosong, maka otomatis akan hilang.

3. Pengklasifikasian Pengguna Data

Matriks yang dihasilkan tersebut memiliki ukuran yang cukup besar. Sehingga untuk menyesuaikan dengan tujuan segmentasi yang hendak dicapai, maka dilakukan pengklasifikasian pada pengguna data, dalam hal ini IP, berdasarkan negara asal pemilik IP. Negara yang dimunculkan pada klasifikasi ini diambil berdasarkan 9 (sembilan) negara yang memiliki frekuensi tertinggi mengakses situs web BPS di Bulan November 2011. Negara-negara yang memiliki frekuensi kecil masuk ke dalam klasifikasi lainnya. Seperti yang terlihat pada Gambar 2.

Gambar 2 di atas didapat dari statistik web pada situs web BPS pada bulan November 2011. Negara dengan kategori unknown adalah alamat IP yang tidak dapat ditelusuri asal negaranya, ada beberapa IP berbayar yang dirahasiakan kepemilikannya, atau yang dikenal dengan private IP number, dan dimasukkan ke dalam klasifikasi lainnya. Sehingga alamat IP berdasarkan asal negara terbagi menjadi 10 (sepuluh) klasifikasi yang dapat dilihat pada Tabel 7.

Berdasarkan klasifikasi tersebut maka seluruh alamat IP yang ada pada tabel kontingensi yang sudah didapat pada Tabel 6 ditransformasi berdasarkan asal negaranya. Untuk mengklasifikasikan alamat IP digunakan program yang dapat dilihat pada Gambar 4 melalui fasilitas tombol "Country Class". Database asal negara diperoleh dari situs web tentang lokasi alamat IP, yaitu http://www.ipaddresslocation.org. Dari transformasi tersebut didapat rekapitulasi akses web BPS berdasarkan negara pada Tabel 8.

Tabel 8 menunjukkan banyaknya akses setiap negara ke web BPS, terlihat bahwa banyaknya akses dari dalam negeri (kode negara 1) sebesar 40,42%. Sedangkan banyaknya akses dari luar negeri (kode

Tabel 8. Rekap Negara yang Mengakses Web BPS

Kode Negara	Frekuensi Akses	Persentase
1	24965	40,42
2	947	1,53
3	11571	18,74
4	4533	7,34
5	3346	5,42
6	626	1,01
7	991	1,60
8	1746	2,83
9	2019	3,27
10	11015	17,84
Total	61759	100

Tabel 9. Tabel Kontingensi Setelah Diklasifikasikan Berdasarkan Negara

Kode	H11	H14	H1	H10	H2	H9	H12	H5	H13	H4	H3	H6	H15	H8
1	21307	1049	1383	771	197	57	12	68	93	17	5	6	0	0
2	624	140	126	9	26	12	0	3	0	2	4	0	1	0
3	931	4228	2486	2760	461	175	333	134	24	16	17	3	2	1
4	374	2579	1183	357	26	9	0	3	0	1	1	0	0	0
5	259	1804	961	109	88	64	7	19	3	5	9	17	1	0
6	115	336	116	28	17	6	0	4	1	2	1	0	0	0
7	687	117	99	15	51	8	0	3	0	6	3	2	0	0
8	574	298	563	204	49	43	0	13	0	2	0	0	0	0
9	439	471	928	84	44	36	0	10	0	2	5	0	0	0
10	4571	2427	3085	249	261	145	143	104	4	12	14	0	0	0

negara 2 s/d 10) sebesar 59,58%. Akses luar negeri paling banyak berasal dari negara Amerika Serikat (kode 3) sebesar 18,74%. Hasil transformasi dari tabel frekuensi yang diklasifikasikan menurut negara dapat dilihat pada Tabel 9.

Perhatikan Tabel 9, kolom Kode menunjukkan kode negara. Sedangkan kolom H11 s/d H8 menunjukkan halaman yang diakses.

4. Penerapan Analisis Korespondensi

Analisis korespondensi dapat digunakan untuk mengetahui kedekatan hubungan antar kategori dari 2 (dua) variabel. Berdasarkan Tabel 9 terdapat 2 (dua) variabel yang dianalisis yaitu negara yang mengakses (Kode) dan halaman yang diakses (H1 s/d H15). Format data pada Tabel 9 diubah terlebih dahulu ke dalam format data yang sesuai dengan *software* statistik seperti pada Tabel 10.

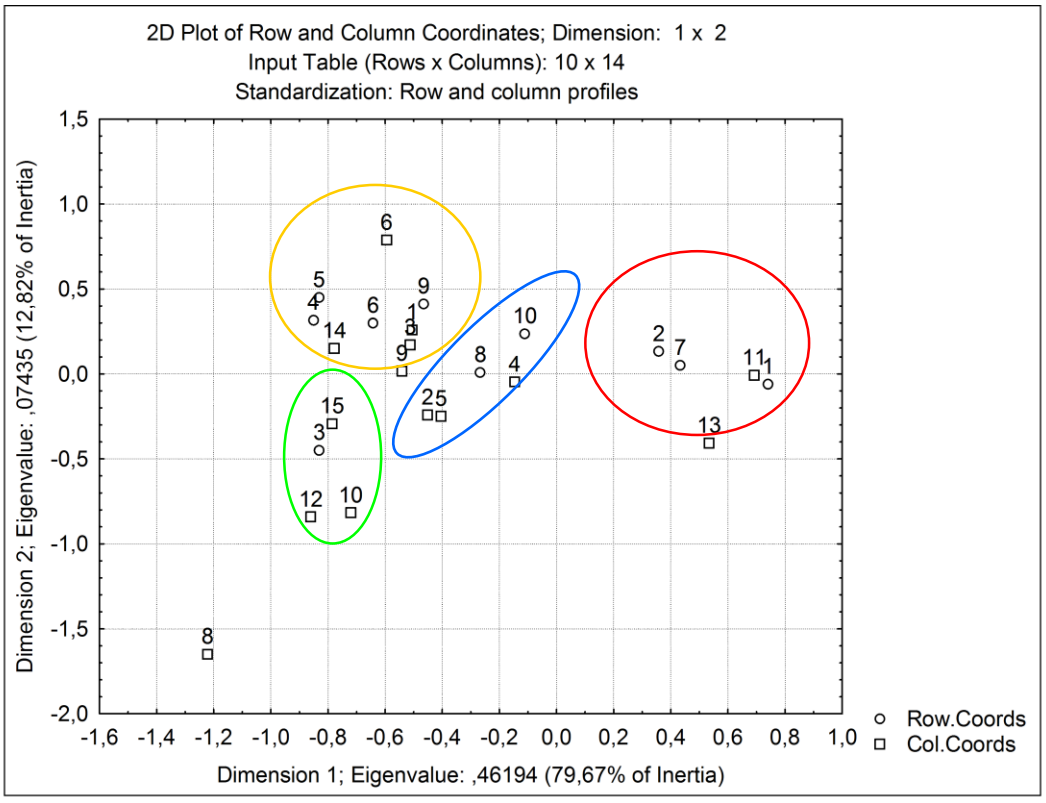
Data di atas kemudian diolah dengan *software* statistik menggunakan analisis korespondensi. *Output* pengolahan dari *software* statistik ditunjukkan pada Gambar 3.

Pada Gambar 3 menunjukkan grafik *symetric map* 2 (dua) dimensi dengan koordinat baris (row coordinates) adalah kode negara dan koordinat kolom (colomn coordinates) adalah kode halaman yang diakses. Sesuai dengan salah satu tujuan analisis korespondensi, terlihat pengelompokkan yang dapat diambil pada grafik. Pengambilan kelompok diambil dengan melihat jarak yang terdekat diantara koordinat tersebut secara subyektif.

Pertama, dapat dilihat pengelompokkan dengan batas merah, negara dengan kode 1, 2, dan 7 (Indonesia, Australia dan Jerman) berkelompok dengan halaman 11 dan 13 (Unduh dan Info Lelang). Sehingga dapat digambarkan

Tabel 10. Format Data Menurut Frekuensi

Klasifikasi	Halaman yang Diakses	Frekuensi
1	11	21307
1	14	1049
…	…	…
2	8	0
3	11	931
3	1	2486
3	10	2760



Gambar 3. *Output Software* Statistik untuk Analisis Koresponden

negara-negara tersebut mengakses halaman-halaman tersebut yang paling banyak. Bukan berarti negara lain tidak mengakses halaman tersebut ataupun bukan berarti negara tersebut tidak mengakses halaman lainnya. Jika dilihat pada pengelompokkan ini, korespondensi antara negara kode 1 (Indonesia) dengan halaman kode 11 (Unduh) sangat dekat.

Kedua, pengelompokkan terjadi pada negara dengan kode 8 dan 10 (Jepang dan Lainnya) dengan halaman berkode 2, 5, dan 4 (Tentang BPS, Pusat Layanan dan Istilah Statistik). Sehingga dapat digambarkan bahwa negara-negara ini berkorespondensi dengan halaman-halaman tersebut. Dapat juga dikatakan negara-negara ini tertarik untuk mengakses data pada BPS melalui bentuk selain web. Misalnya melalui perpustakaan maupun konsultasi statistik yang ada dalam Pusat Layanan.

Ketiga, pengelompokkan terjadi pada negara dengan kode 4, 5, 6 dan 9 (Inggris, Cina, Malaysia dan Singapura) dengan halaman berkode 1, 3, 9, dan 14 (Beranda, Rencana Strategis BPS, Publikasi BPS, dan Subyek Satistik) yang menggambarkan bahwa negara-negara ini mengakses dengan hampir merata terhadap halaman-halaman yang ada pada situs web BPS, tetapi paling berkorespondensi dengan halaman-hlaman tersebut. Jika dilihat pada pengelompokkan ini, negara dengan kode 4 (Inggris) sangat dekat korepondensinya halaman berkode 14

(Subyek Statistik) yang menggambarkan tingginya akses negara tersebut dalam membuka halaman Subyek Statistik yang berisi tabel-tabel statistik berdasarkan subyek.

Selanjutnya adalah pengelompokkan antara negara dengan kode 3 (USA) dan halaman dengan kode 10, 12 dan 15 (Berita Resmi BPS, Berita dan Website BPS Provinsi). Hal ini menggambarkan negara tersebut paling banyak mengakses halaman-halaman tersebut. Sedangkan pada halaman berkode 8 (Sekolah Tinggi Ilmu Statistik) berada pada posisi yang jauh dari kelompok manapun. Hal ini menggambarkan halaman ini yang paling jarang diakses oleh negara-negara tersebut.

KESIMPULAN DAN SARAN

Berdasarkan hasil dan pembahasan, dapat diambil beberapa kesimpulan bahwa pengguna situs web BPS yang diwakili oleh alamat IP dapat dikelompokkan dengan halaman yang diakses berdasarkan asal negara, sehingga didapat segmentasi pengguna data situs web BPS. Secara garis besar menjadi 3 (tiga) kelompok:

1. Berdasarkan Gambar 4 terjadi pengelompokkan pada negara Indonesia, Australia dan Jerman terhadap halaman "Unduh" dan "Info Lelang", bahkan korespondensi antara Indonesia dan halaman "Unduh" sangat dekat. Ini bisa diartikan bahwa yang mengunduh halaman web BPS paling banyak berasal dari Indonesia. Sedangkan untuk Australia dan Jerman juga banyak mengakses unduh dengan jarak yang hampir sama dengan Indonesia mengakses Info Lelang.
2. Kedua, pengelompokkan terjadi pada negara Jepang dan Lainnya dengan halaman "Tentang BPS", "Pusat Layanan" dan "Istilah Statistik". Sehingga dapat digambarkan bahwa negara-negara ini berkorespondensi dengan halaman-halaman tersebut. Dapat juga dikatakan negara-negara ini tertarik untuk mengakses data pada BPS melalui bentuk selain web. Misalnya melalui perpustakaan maupun konsultasi statistik yang ada dalam Pusat Layanan.
3. Kelompok ketiga terjadi pada negara Inggris, Cina, Malaysia dan Singapura dengan halaman "Beranda", "Rencana Strategis BPS", "Publikasi BPS", dan "Subyek Satistik". Negara Inggris sangat dekat korepondensinya halaman "Subyek Statistik" yang menggambarkan tingginya akses negara tersebut dalam membuka halaman Subyek Statistik yang berisi tabel-tabel statistik berdasarkan subyek.

Kesimpulan yang diwakili oleh ketiga kelompok ini, bukan berarti negara-negara tersebut tidak mengakses halaman-halaman lainnya. Secara korespondensi bisa dilihat kedekatan yang paling sering diakses. Yang menarik adalah halaman berkode 8 (Sekolah Tinggi Ilmu Statistik) yang berada jauh dari kelompok manapun, hal ini bisa dipelajari lebih lanjut.

Berdasarkan kesimpulan tersebut, maka penulis menyarankan beberapa hal, sebagai berikut:

1. Hasil pengelompokkan dapat digunakan sebagai bahan pertimbangan dalam mengembangkan situs web BPS, berhubungan dengan tampilan dan kemudahan akses dalam membuka halaman-halaman yang sering diakses.
2. Penyempurnaan untuk halaman web berbahasa asing, berhubungan dengan eratnya korespondensi negara luar dalam mengakses halaman-halaman yang ada pada situs web BPS.

DAFTAR PUSTAKA

Almind and Ingwersen. 1997. Informetric analyses on the World Wide Web: Methodological Approaches to Webometrics. E-Journal on-line Melalui http://www.cindoc.csic.es/cybermetrics/

Bjorneborn and Ingwersen. 2004. Toward a Basic Framework for Webometrics. E-Journal on-line Melalui http://www.interscience.wiley.com/cgi-bin/abstract/109594194/ABSTRACT

BPS. 2011. Laporan Reformasi Birokrasi Badan Pusat Statistik. Jakarta: BPS.

Cox, et al. 2001. *Multidimensional Scalling* (Second Ed.). New York: CRC Press LCC. E-book.

Greenacre, J. 1984. *Theory and Application of Correspondence Analysis*. London: Academic Press. E-book.

Izenman, A.J. 2008. *Modern Multivariate Statistical Techniques*. New York: Springer. E-book.

Khodra, M.L. 2003. Text Mining Kategori Teks Naive Bayes. E-Journal on-line Melalui http://kur2003.if.itb.ac.id/file/TextMiningKlasifikasiNB.pdf

Nicholson, S. 2006. The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services. E-Journal on-line Melalui http://arizona.openrepository.com/arizona/bitstream/10150/106175/1/nicholson2.pdf

Santoso, B. 2007. *Data Mining Teknis Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu.

Srivastava, et al. 2000. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. E-Journal on-line Melalui http://nlp.uned.es/WebMining/Tema5.Uso/srivastava2000.pdf

Supriyadi, Y. 2011. *Aplikasi Teknik Bibliometric pada Analisis Data Paten*. Seminar Statistik Nasional 2011.

Tarapanoff, K, et al. 2001. *Intellegence Obtained by Applying Data Mining to a Database of French Theses on The Subject of Brazil* . Information Research, Vol. 7 No. 1, October 2001.

Thelwall, M. 2007. Bibliometrics to Webometrics. E-Journal on-line Melalui http:// www.scit.wlv.ac.uk/ ~cm1993/ papers/ JIS-0642-v4-Bibliometrics-to- Webometrics.pdf

Thelwall, M. 2009. Introduction to Webometrics: Quantitative Web Research for the Social Sciences. E-Journal on-line Melalui http://www.morganclaypool.com/doi/abs/10.2200/S001-76ED1V01Y200903ICR004

Web Mining. Melalui http://en.wikipedia.org/wiki/Web_mining

Xu, et al. 2011. *Web Mining and Social Networking*. New York: Springer. E-book.