Aplikasi Pemodelan Bahasa Secara Statistik dalam Bisnis Periklanan di Internet

Nuansa Lembayung / 1820928
Program Studi Sistem dan Teknologi Informasi
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]18209028@std.stei.itb.ac.id

Makalah ini membahas tentang penerapan teori-teori dalam pemodelan bahasa secara statistik yang banyak digunakan dalam penentuan iklan kontekstual yang sekarang berkembang di internet. Dalam penentuan iklan kontekstual, digunakan metode pemodelan bahasa dengan statistik sehingga iklan yang ditayangkan sesuai dengan konteks yang sedang dibaca atau dilihat oleh pengakses suatu halaman web. Dalam pemodelan bahasa secara statistik, perhitungan menggunakan teori proses Markov untuk menghitung probabilitas dalam meramalkan bahasa yang sesuai konteksnya berdasarkan daftar bahasa yang ada dan diambil dengan pendekatan statistika. Hasil yang ingin didapat adalah kita dapat menentukan iklan apa yang paling tepat dimunculkan dalam konteks sebuah halaman tertentu di internet sehingga iklan yang dimunculkan dapat tepat sasaran dan sesuai dengan minat dari pengakses halaman tersebut sehingga iklan yang ditampilkan dapat menjadi efektif dan sesuai dengan target.

Kata kunci: *iklan kontekstual, pemodelan bahasa, proses Markov, statistik.*

I. PENDAHULUAN

Pada saat ini, pengguna internet bertumbuh demikian pesatnya, hampir setiap negara di dunia ini punya akses ke jaringan internet. Dari miliaran pengguna internet di seluruh dunia yang akan terus bertambah, terdapat triliunan halaman website yang ada, dan dari triliunan halaman tersebut, tidak terhitung jumlah kata-kata yang ada di seluruh internet yang pastinya ada dalam berbagai bahasa di dunia.

Di sisi lain, internet tidak dapat dipungkiri menjadi media yang sangat efektif untuk menyebarkan informasi pada saat ini dan juga di masa yang akan datang. Hal ini menyebabkan bisnis di bidang ini pun berkembangan dengan sangat pesat. Begitu pula dengan bisnis periklanan. Periklanan lewat media ini menjadi begitu berkembang dan dianggap efektif karena penyebarannya yang cepat, namun dengan biaya yang relatif lebih murah daripada beriklan di media cetak maupun media elektronik konvensional.

Namun, pada perkembangannya, konsep periklanan di internet dianggap tidak ada bedanya dengan konsep iklan media konvensional yang hanya berganti medium menjadi media internet tanpa ada suatu revolusi dalam beriklan. Padahal, internet seharusnya memunyai potensi yang jauh lebih besar daripada media konvensional karena kemudahan akses dan keterbukaannya.

Lalu timbul suatu gagasan baru yang revolusioner tentang iklan berbasis kontekstual. Suatu iklan yang muncul sesuai dengan konteks dari suatu halaman internet yang dibuka. Sebagai contoh jika terdapat halaman yang membahas tentang musik, lebih spesifik, musik rock, maka contoh iklan yang paling pas adalah iklan gitar elektik contohnya. Beda halnya dengan contohnya sama-sama membahas musik, namun lebih spesifik halaman tentang musik jazz, tentunya iklan gitar elektrik sudah kurang relevan lagi walaupun masih ada hubungannya. Maka untuk kasus tersebut iklan yang dianggap paling pas adalah iklan gitar akustik atau iklan saxophone contohnya. Tentunya hal ini akan meningkatkan efektifitas iklan menjadi sangat tinggi, dan hal inilah yang dilakukan oleh perusahaan periklanan pada saat ini di internet.

Tentunya, dalam kasus-kasus di atas, perlu dilakukan pendataan yang sangat rumit jika ingin menghasilkan kata-kata yang tepat untuk iklan tertentu. Tidak hanya itu, mendata kata-kata satu per satu tentunya juga akan sangat sulit dan memerlukan waktu yang lama sedangkan segalanya dituntut untuk dilakukan secara cepat.

Dari situlah pemodelan bahasa secara statistik berguna. Dengan cara pemodelan ini, kata-kata yang sudah ada diambil sampelnya dan dengan metode proses Markov, dimana proses Markov akan berguna untuk mengkalkulasikan peluang dari kata-kata yang muncul terhadap sebuah rangkaian kata. Dari sini akan ditaksir kata-kata apa saja yang relevan terhadap kata-kata yang ada di sebuah halaman web yang terlebih dahulu diindeks kata-kata yang sering muncul mulai dari kueri yang dimasukkan pengguna, judul, sampai isi dari halaman internet tersebut. Pengindeksan dilakukan terlebih dahulu sehingga dalam mencocokan kata-kata yang muncul, sistem sudah memunyai semacam acuan dalam menentukan probabilitasnya. Acuan ini dipercaya oleh sistem sebagai acuan yang baik. Semakin banyak acuan yang dimiliki oleh sistem, tentunya hasil dari probabilitas kueri yang didapat akan semakin akurat dan baik.

Dengan penggunaan metode tersebut, diharapkan dari serangkaian kata yang disiapkan untuk sebuah iklan dapat dimunculkan di saat dan tempat yang tepat sehingga

kemunculan iklan akan sesuai dengan konteks dari halaman web tersebut.

II. METODE

II.1 Teori Proses Markov

Proses Markov adalah sebuah teori yang dikeluarkan oleh Andrey Markov yang menyatakan bahwa dalam proses stokastik, pemodelan matematik untuk menghitung peluang keadaan dari proses pada waktu yang akan datang tidak dipengaruhi oleh keadaan pada masa lampau karena keadaan dari masa lampau dan masa depan independen dan tidak ada hubungannya. Masa lampau sudah lewat dan sudah tidak berpengaruh lagi terhadap kondisi masa depan. Keadaan pada masa yang akan datang hanya dipengaruhi oleh kondisi saat ini . Kejadian atau kondisi saat ini dapat menjadi acuan untuk menentukan kemungkinan kejadian berikutnya. Kejadian yang digunakan dalam penentuan kejadian masa depan disebut sebagai "order" yang jika mengambil satu sampel kondisi terakhir disebut "first order process" dan jika mengambil dua sampel disebut "second order process" dan begitupula seterusnya. Dan yang harus ditekankan bahwa sampel tersebut diambil pada keadaan saat itu.

Suatu proses X pada waktu t, merupakan proses Markov jika untuk setiap n dan untuk setiap

$$\{X_t, t \geq 0\}$$
$$t_1 < t_2 < t_3 < \cdots < t_n < t_{n+1}$$

Untuk kondisi di atas, berlaku kaitan:

$$P\,(X_{n+1}|X_n, X_{n-1}, \dots, X_1) = P(X_{n+1}|X_n)$$

Rumus di atas merupakan rumus peluang berkait dimana proses selanjutnya berkaitan dengan proses sebelumnya. Maka dari itu, untuk menghitung peluang yang akan datang pada X_{n+1}, kita harus mengetahui kondisi atau peluang pada masa kini (X_n) yang terjadi di waktu (t_n) saat itu. Hal ini menunjukkan bahwa proses Markov merupakan proses yang sama sekali tidak bergantung kepada kondisi di masa lalu, melainkan hanya kondisi saat itu saja.

Selanjutnya jika dimisalkan X_n memunyai ruang sampel berupa himpunan cacah berhingga, maka proses Markov disebut sebagai rantai Markov.

Contoh dari dari rantai Markov ini adalah barisan bilangan bulat saling bebas dan memunyai distribusi peluang yang sama sehingga jika X_n adalah sebuah bagian sebuah himpunan proses, maka S_n merupakan jumlah dari seluruh himpunan tersebut seperti pada rumus di bawah ini:

$$S_n = \sum_{i=1}^{n} X_i$$

Sehingga jika dikaitkan dengan rumus peluang sebelumnya, dimana proses-proses tersebut kita ganti menjadi kumpulan proses, rumus:

$$\begin{aligned} P\,(S_{n+1}|S_n, S_{n-1}, \dots, S_1) &= P\Big(\sum_{i=1}^{n+1} X_i \,\Big|\, X_1, \dots, \sum_{i=1}^{n-1} X_i, \sum_{i=1}^{n} X_i\Big) \\ &= P\Big(\sum_{i=1}^{n+1} X_i = S_{n+1} \,\Big|\, \sum_{i=1}^{n} X_i = S_n\Big) \end{aligned}$$

Dari rumus di atas dapat terlihat bahwa kemungkinan-kemungkinan pada n sebelumnya akhirnya akan tetap menjadi satu kemungkinan pada saat ini yang memengaruhi peluang pada n+1 dan rumus untuk kumpulan proses juga berlaku di sini.

Dari teori ini dapat dikembangkan menjadi berbagai teori lainnya seperti menghitung proses acak pada gerak Brown, dll.

II.2 Pemodelan Bahasa

Pemodelan bahasa (language modelling) merupakan sebuah metode cara penaksiran dari kata-kata sebuah bahasa. Kata-kata yang ditaksir dalam pemodelan bahasa dibuat sebuah distribusi kemungkinan yang akan diproses untuk berbagai keperluan. Distribusi itu akan menunjukkan penaksiran yang paling tepat untuk rangkaian kata tertentu.

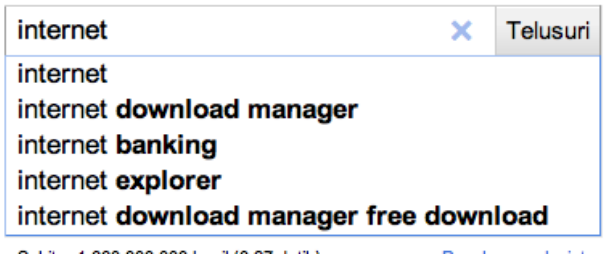
Bahasa yang biasa digunakan manusia mengandung berbagai ketidakteraturan terlepas dari aturan-aturan bahasa yang ada. Ketidakteraturan ini yang membuat komputer sangat terbatas dalam menentukan apa saja maksud yang ingin disampaikan jika pengguna memasukkan beberapa kata kunci.

Di sinilah pemodelan bahasa digunakan, pemodelan bahasa digunakan untuk menentukan informasi apa saja yang bisa didapat pada saat ini untuk digunakan sebagai acuan dalam menentukan bahasa selanjutnya yang dimaksudkan oleh pengguna.

Aplikasi penggunaan pemodelan bahasa sangat beragam karena pada dasarnya pemodelan bahasa ini merupakan pengkoreksi dari segala bahasa manusia sehingga mesin dapat mengenalinya, memprosesnya dan bahkan memerbaikinya. Beberapa aplikasi penggunaan pemodelan bahasa contohnya digunakan dalam pemrosesan bahasa seperti pengenal perkataan yang biasa terdapat di alat-alat elektronik seperti handphone. Juga terdapat di alat untuk menerjemahkan dan penerima informasi digital.

Sebagai contoh pada saat menggunakan mesin pencari di internet, kita sering kali diberikan kemungkinan-kemungkinan pencarian dari kata-kata yang dimasukkan

seperti pada gambar 2.1.



Gambar 2.1 Contoh aplikasi pemodelan bahasa

Perhitungan dalam pemodelan bahasa banyak menggunakan dasar dari teori Markov karena untuk memprediksi rangkaian kata contohnya pada proses penerjemahan, dibutuhkan pemodelan pada rangkaian kata sebelumnya untuk menentukan terjemahan rangkaian kata sesudahnya. Dengan menggunakan teori proses Markov, kata yang dimasukkan pada saat sebelumnya dibaca dan dikalkulasi dengan menggunakan pendekatan statistik dicari relevansinya dengan kata-kata sejenis pada dokumen lain di internet dan kecenderungannya sehingga menghasilkan kueri yang sesuai dan relevan dengan yang sebelumnya telah dimasukkan.

II.4 Mesin Penerjemah Statistik

Mesin penerjemah statistik adalah mesin penerjemah kalimat yang menggunakan konsep probabilitas. Mesin penerjemah statistik merupakan konsep pemotongan informasi (parsing) dan melakukan perbadingan terhadap segmen-segmen di dalam sebuah dokumen yang dapat berupa kata maupun rangkaian kata.

Sebagai contoh, jika terdapat dua kalimat A dan B dalam sebuah dokumen atau halaman web:

A: Perkembangan teknologi digital sangat pesat
B: Internet sebagai sarana utama perkembangan teknologi

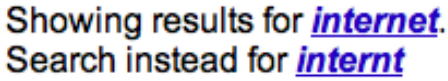
Maka mesin penerjemah statistik akan memisahkan dua kalimat tersebut dan mengkalkulasi peluang kemunculan kata dengan menggunakan rumus peluang berkait seperti di bawah ini:

$$P(B|A) = \frac{P(A) \cap P(B)}{P(A)}$$

Mesin ini akan menganalisa kata-kata apa saja yang berkaitan di dua kalimat di atas dan menghasilkan suatu probabilitas kemungkinan relevansi antara kata-kata di kedua kalimat tersebut. Pada contoh di atas, mesin akan mendeteksi bahwa terdapat relevansi antara kata "perkembangan teknologi" dengan "internet".

Statistik ini dilakukan berjuta-juta kali terhadap berbagai bahasa dan tulisan di internet sehingga menghasilkan analisis kemungkinan yang akurat yang akan dijadikan referensi untuk pencarian selanjutnya.

Namun tidak hanya di situ, dalam perkembangan yang lebih canggih lagi, analisis ini tidak hanya dilakukan terhadap kata per kata, tetapi juga sudah mencakup kalimat per kalimat sebagai sebuah kesatuan sehingga mesin dapat mendeteksi kalimat-kalimat yang relevan. Dari situ mesin dapat mengkoreksi masukan kalimat pengguna yang salah maupun mencari hal-hal yang berkaitan dengan kalimat tersebut. Begitupula dengan kata dimana mesin pencari sering mendeteksi kesalahan dalam penulisan kata seperti pada gambar 2.2.



Gambar 2.2 Contoh pembenaran dalam mesin pencari

Terlihat bahwa mesin pencari telah melakukan pengindeksan secara statistik terhadap kata "internet" sehingga jika pengguna memasukkan huruf-huruf yang menurut data statistik probabilitasnya merujuk kepada kata "internet", maka kesalahan penulisan "internt" akan diberikan peringatan karena intensi penggunaan kata internet sudah sangat banyak yang menghasilkan probabilitas yang tinggi pula.

II.4 Penerapan Proses Markov pada Pemodelan Bahasa Secara Statistik

Seperti yang telah dijelaskan sebelumnya, pemodelan bahasa merupakan aplikasi statistik yang menggunakan teori Markov. Pada kasus ini, yang ingin dijadikan acuan dan didata statistiknya oleh mesin penerjemah statistik adalah berupa sebuah bahasa, bahasa manusia.

Seperti yang sudah dijelaskan pada teori Markov (lihat poin II.1), jika kita mengasumsikan titik sampel X_n sebagai sebuah kata (W), dan S_n yang merupakan himpunan dari X_n adalah sebuah kumpulan dari kata-kata yang bisa berupa kalimat atau pun kata-kata acak yang dikumpulkan. Sehingga:

$$S = W_1, W_2, \dots, W_n$$

Dan jika k merupakan banyaknya kata-kata yang dijadikan model untuk menentukan probabilitasnya, maka S merupakan proses Markov dengan rumus probabilitas:

$$P_n(S) = \prod_{i=1}^{k} P\ (w_1 | w_{i-1}, w_{i-2}, \dots, w_{i-n+1})$$

Jika n = 1, model yang dipakai adalah *unigram language model* yang merupakan proses Markov dalam menentukan probabilitas dari kata-kata secara individual yang berarti kata per kata dikalkulasi probabilitasnya sendiri-sendiri sehingga dalam konteks mesin pencari, kasus yang akan muncul seperti perbaikan kata yang salah pada saat penulisan karena statistik yang dicatat adalah hanya kata per kata. Pada model ini, urutan kata pada

kalimat tidak dihitung dan dikalkulasi. Dalam konteks iklan di internet, suatu iklan didasarkan kepada kata per kata yang muncul di halaman itu sehingga di kasus ini iklan yang muncul hanya melihat intensitas dari kata tersebut dan tidak melihat konteks pembicaraan tulisan tersebut secara garis besar.

Lain halnya jika n = 2, model yang dipakai adalah *biagram language model* yang merupakan proses Markov dalam menentukan probabilitas dari informasi kata-kata yang bersama-sama sehingga sudah bisa mendeteksi kesalahan pada kalimat. Semakin tinggi tingkat n-nya, semakin komplek penentuan probabilitas dari sekumpulan kata-kata dan semakin akurat hasilnya. Dalam konteks iklan internet, iklan ini tidak hanya melihat dari kata per kata, tetapi juga sudah bisa melihat dari kalimat per kalimat dan konteks apa yang dibawa tulisan tersebut yang akan disesuaikan dengan iklan yang muncul.

Dalam dunia internet, sistem sudah memiliki gambaran dokumen-dokumen acuan yang sangat besar sebagai acuan referensi mereka dalam menentukan probabilitas dari sekumpulan dokumen (halaman web) lain yang dianalisis.

Dari dokumen (halaman web) lain yang ditemukan di internet, maka dicari padanannya dari dokumen acuan yang sudah ada di database besar dalam pemodelan bahasa ini. Sistem akan memperkirakan dengan menggunakan rumus:

$$\arg\max P(D|Q) = \arg\max P(Q|D)P(D)$$

Dimana Q adalah kueri atau kata-kata yang dicari referensinya (probabilitasnya terhadap kata lain) dan D adalah dokumen. Dengan begini, kita akan menghitung sebaran dari kata-kata tersebut di setiap dokumen dan menghitung probabilitas kueri sampel dari setiap dokumen dan dokumen akan diurutkan sesuai nilai probabilitasnya.

Dalam artikel yang ditulis Ponte dan Croft, mereka menggunakan metode Bernoulli multivariat untuk menghitung P(Q|D) yang menganggap sebuah kueri merupakan sebuat vektor dengan atribut biner yang masing-masing atributnya menyimpan sebuah istilah unik dalam kosakata indeks dan menandakan apakah istilah tersebut ada atau tidak dalam kueri. Jumlah kemunculan dari kueri tersebut tidak diperhitungkan, namun asumsi yang mendasari model ini adalah bahwa semua atribut bernilai biner yang berarti 1 jika mewakili sebuah istilah dan 0 jika tidak. Sehingga P(Q|D) dapat dirumuskan dari dua probabilitas yaitu kemungkinan kemunculan dan kemungkinan ketidak-munculan istilah tersebut. Yang dalam rumusberupa:

$$P(Q|D) = \prod_{w \in Q} P(w|D) \prod_{w \in Q} (1.0 - O(w|D))$$

Pada rumus di atas, P(w|D) dihitung dengan memanfaatkan probabilitas dari rata-rata kemunculan w di dalam dokumen D yang mengandungnya. Untuk frekuensi dokumen yang memunculkan kueri tersebut juga diperhitungkan walaupun tidak digunakan secara menyeluruh.

Namun, dalam penerapannya, terkadang dokumen-dokumen yang dijadikan referensi kurang relevan terhadap kueri yang dicari. Sehingga jika kita mengambil sebuah kueri dalam dokumen R yang belum ada pembandingnya, maka dengan menggunakan probabilitas bersama dapat dihasilkan rumus:

$$P(w|R) \approx P(w|Q) = \frac{P(w, q_1, \dots, q_m)}{\sum_{v \in vocabulary} P(v, q_1, \dots, q_m)}$$

Dari situ baru bisa ditentukan probabilitas dari kueri-kueri yang ada sekalipun di dalam database mesin tidak ada referensi yang tetap, namun dapat dicari dengan menggunakan padanan dari kamus yang ada di bahasa terkait.

Dalam penerapannya, sebuah layanan penyedia iklan di internet akan memakai proses penentuan probabilitas kata-kata ini sebagai acuan dari dokumen (halaman web) yang ditampilkan seberapa kuat relevansinya dan kata-kata apa saja yang biasa muncul di dokumen dengan kata-kata tertentu sehingga dalam menentukan iklan apa, mesin statistik dapat menentukan "kebiasaan" dari sebuah kueri dan mengambil iklan dengan probabilitas tertinggi dari kueri-kueri yang relevan.

Tidak hanya di situ, dengan menggunakan *bidiagram language model* kueri yang didapat tidak hanya dicocokan dengan kata-kata yang sudah ada, tetapi juga dengan keseluruhan kalimat dan semua konteks pembicaraan dari sebuah tulisan. Degan begitu, ketepatan dari penampilan iklan di internet menjadi jauh lebih efektif karena pada dasarnya seseorang yang membuka sebuah halaman internet tentang suatu topik akan jauh lebih mudah mengakses iklan sesuai dengan topik yang sedang dibaca dan memiliki kemiripan hubungan tertentu.

V. KESIMPULAN

- Iklan internet merupakan iklan modern yang berbasis kontekstual dimana isi dari iklan tersebut berdasarkan dari isi halaman web itu sendiri
- Segala hal yang berhubungan dengan pengenalan bahasa oleh mesin dapat dikerjakan dengan pemodelan bahasa
- Pemodelan bahasa dilakukan dengan menggunakan mesin penerjemah statistik
- Mesin penerjemah statistik menggunakan dasar teori proses Markov dalam penerapannya
- Teori proses Markov menggunakan pendekatan probabilistik untuk menentukan kemungkinan kemunculan sebuah kata atau kalimat berdasarkan kepada data statistik yang sudah ada
- Teori proses Markov dalam penerapannya di pemodelan bahasa dapat menggunakan *unigram language model* yang menentukan probabilitas kata-

kata secara individual dan *n-gram language* model yang menentukan probabilitas dari rangkaian kata-kata dengan konteks yang lebih luas

REFERENSI

[1] http://iperpin.wordpress.com; Tanggal Akses: 14/12/2010

[2] http://en.wikipedia.org/wiki/Language_model; Tanggal Akses: 14/12/2010

[3] http://en.wikipedia.org/wiki/Markov_chain; Tanggal Akses: 14/12/2010

[4] http://docs.google.com/viewer?a=v&q=cache:7xekqhGbPCMJ:achmad.blog.undip.ac.id/files/2009/06/markov.pdf+proses+markov&hl=en&pid=bl&srcid=ADGEESjLrUBNe98Cf-u9iNdZLKoc9nJGUYZjozmmKeIaEol76WiEtd04Q7ebIJVKM61r564jmQTd9LHAH_ei4KocaaF3LiSEdtpWuFv7Cn8Q7yf-1EmsN7BtgkaLbXR4vCLvb22rqqIp&sig=AHIEtbSS8iXBjU2T98oVQFkr5vJC3VjeIw; Tanggal Akses: 15/12/2010

[5] http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html; Tanggal Akses: 15/12/2010

[6] T. Hansel, "PENERJEMAHAN DOKUMEN INGGRIS-INDONESIA MENGGUNAKAN MESIN PENERJEMAH STATISTIK DENGAN WORD REORDERING DAN PHRASE REORDERING" unpublished.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 16 Desember 2010

ttd
Nuansa Lembayung
18209028