ANCANGAN PENYUSUNAN TESAURUS

Zainal A. Hasibuan
Fakultas Ilmu Komputer - Universitas Indonesia

1. Latar Belakang

Sudah lama orang menyadari bahwa penyusunan buku-buku, artikel-artikel, atau dokumen lainnya, ke dalam kelas-kelas tertentu, akan sangat membantu untuk menemukan kembali (*retrieval*) dokumen tersebut bagi orang yang membutuhkannya. Oleh karena itu lahirlah berbagai macam petunjuk bagi pencari informasi. Petunjuk tersebut dapat berupa sistem klasifikasi penyimpanan dokumen seperti: *Colon Classification*, *Library of Congress Classification*, *Dewey Decimal Classification*, *Universal Decimal Classification* dan lain sebagainya, yang sering kita jumpai di perpustakaan-perpustakaan. Di samping berbagai macam klasifikasi tersebut, orang juga memerlukan petunjuk untuk dapat memperoleh isi atau konsep yang terkandung di dalam dokumen-dokumen tersebut. Maka lahirlah sistem pengindeksan (*indexing system*). Sebagai suatu contoh, pada bagian akhir suatu buku, selalu dicantumkan kata-kata indeks (*index terms*). Kata-kata indeks ini selalu kita pergunakan kalau kita hendak mencari suatu topik tertentu yang terkandung dalam buku tersebut.

Dengan berkembangnya ilmu pengetahuan, dan semakin banyaknya dokumen tertulis yang hendak disimpan, maka petunjuk yang telah ada terasa semakin tidak cukup untuk bisa membantu orang menemukan informasi yang dibutuhkannya. Sementara itu sistem klasifikasi dan sistem pengindeksan yang telah ada, semakin rumit untuk bisa mengakomodasi berbagai macam cabang ilmu pengetahuan dan berbagai macam topik yang bisa diwakilkannya. Maka lahirlah tesaurus, yang bertujuan untuk mengkoordinasikan kata-kata indeks dari sekelompok dokumen, untuk keperluan pencarian kembali dokumen tersebut (*document retrieval*). Di samping itu, tesaurus dipergunakan juga untuk membimbing si pemakai untuk memperoleh kata-kata yang tepat untuk dapat dipergunakan dalam penulisan. Contohnya, *Roget's Thesaurus*

dirancang untuk membantu seseorang dalam menulis dan memilih kata-kata yang sesuai dengan yang diinginkannya.

Tesaurus berasal dari kata Yunani (*Greek*) yang berarti "*a treasure house*". Yang membedakan tesaurus dengan kamus atau sistem pengindeksan lainnya adalah, tesaurus terdiri dari beberapa kelompok kata-kata yang disusun menurut abjad, di mana pada setiap kelompok, kata-kata tersebut mempunyai hubungan antara satu kata dengan kata lain, maupun kelompok lain. Hubungan tersebut dapat berupa hubungan yang bersifat hirarkis, maupun horizontal. Contoh hubungan yang bersifat hirarkis adalah adanya "*broader term*", yang mengaitkan dengan kata-kata yang lebih umum maknanya, dan "*narrow term*", yang mengaitkan dengan kata-kata yang lebih spesifik maknanya. Sedangkan yang bersifat horizontal misalnya hubungan "*related term*", sinonim, ataupun antonim. Di samping itu, setiap kata bisa juga diberikan catatan cakupannya (*scope note*).

Kata-kata yang tercantum dalam tesaurus berikut hubungannya antara suatu kata dengan kata lainnya, dapat juga dikategorikan sebagai representasi dari suatu basis pengetahuan dari suatu domain ilmu pengetahuan. Hal ini dikarenakan kata-kata yang berada dalam suatu kelompok tersebut, merupakan representasi dari suatu konsep, atau suatu topik dari suatu ilmu pengetahuan.

2. Ancangan Pembuatan Tesaurus

Prosedur pembuatan tesaurus bukanlah sesuatu kegiatan yang kaku. Tesaurus harus bisa berkembang untuk menyesuaikan dirinya sebagai suatu alat yang berfungsi membantu orang mencari informasi terhadap sekelompok dokumen, dan membantu orang memilih kata-kata yang tepat dalam merepresentasikan suatu konsep dalam suatu kegiatan penulisan. Berikut ini, disampaikan langkah-langkah apa yang hendaknya diambil dalam pembuatan suatu tersaurus.

2.1 Pemilihan Kata-kata (*Terms Selection*)

Hal pertama yang perlu ditekankan adalah, tesaurus yang dibuat sebaiknya berorientasi kepada para pemakai. Oleh karena itu, tesaurus tersebut perlu melampirkan keterangan mengenai cakupannya, dan aturan-aturan pemakaiannya. Sebagai contoh, banyak tesaurus yang dibuat hanya terpaut pada disiplin ilmu tertentu. Para pemakai perlu diberitahukan bahwa, kata-kata yang dipakai di dalam tesaurus tersebut, disarikan dari sekelompok artikel ilmiah dari disiplin ilmu yang bersangkutan. Begitu juga halnya kalau yang dibuat tersebut merupakan tesaurus yang bersifat umum.

Pada dasarnya, tidak ada petunjuk yang baku mengenai pemilihan kata-kata apa yang sebaiknya diikutsertakan sebagai entri di dalam suatu tesaurus. Kapan suatu kata layak dicantumkan di dalam suatu tesaurus? Jawaban pertanyaan yang demikian, sebagian besar tergantung kepada pengetahuan dan keputusan si pembuat tesaurus. Keputusan lain yang perlu diambil adalah mengenai ukuran dari suatu tesaurus. Berapa banyak kata-kata yang selayaknya dicantumkan di dalam suatu tesaurus, sehingga tesaurus tersebut bisa secara efektif digunakan untuk membantu si pemakai? Hal inipun tidak ada suatu jawaban yang pasti.

Ukuran suatu tesaurus yang terlalu besar, dapat mempersulit si penyusun tesaurus dalam mengorganisir kata-kata yang hendak dicantumkan. Dari segi pemakai, ukuran tesaurus yang terlalu besar, bisa jadi akan membuat si pemakai kehilangan arah, mengenai konsep atau topik yang hendak dia representasikan. Sebaliknya, ukuran tesaurus yang terlalu kecil, kemungkinan besar tidak akan bisa memberikan pelayanan yang memuaskan kepada para pemakai.

Walaupun jawaban-jawaban dari pertanyaan di atas tidak ada yang definitif, tapi ada beberapa hal yang bisa dijadikan acuan. Hal yang pertama adalah, seperti telah disampaikan sebelumnya, suatu tesaurus sebaiknya berorientasi kepada para pemakai. Oleh karena itu, suatu tesaurus sebaiknya menggunakan perbendaharaan kata-kata para pemakai yang dijadikan sebagai entri, bukan perbendaharaan kata-kata si pembuat tesaurus. Untuk memenuhi hal ini, perbendaharaan kata-kata tersebut sebaiknya disarikan dari tulisan-tulisan yang sudah ada, misalnya dari

artikel atau dokumen. Sumber lain yang dapat dipakai adalah daftar "*subject headings*", dari suatu disiplin. Sehingga orang yang ahli di bidang tersebut, dapat dijadikan sebagai sumber untuk berkonsultasi. Hal yang kedua yang perlu diingat adalah, tesaurus tersebut merupakan suatu alat. Suatu alat harus dirancang sedemikian rupa, agar mudah digunakan oleh orang lain. Perancangan tesaurus akan diuraikan pada bagian berikut ini.

2.2 Rancangan Pengelompokan Kata-kata Tesaurus

Dalam merancang suatu tesaurus, perlu terlebih dahulu membagi domain dari suatu disiplin ilmu ke dalam kelas-kelas tertentu. Kelas-kelas merupakan sub-disiplin subdisiplin dari ilmu pengetahuan tersebut. Pada umumnya, hal ini juga ditentukan secara subjektif oleh si pembuat tesaurus. Sumber yang paling ideal untuk berkonsultasi dalam penentuan kelas-kelas tersebut adalah para pakar dari disiplin ilmu yang bersangkutan.

Tahap berikutnya adalah menentukan istilah-istilah yang mewakili suatu topik, subjek, atau konsep, ke dalam kelas-kelas tersebut. Istilah di sini dapat juga disebut sebagai deskriptor. Deskriptor ini diperoleh dari hasil proses pada tahapan yang telah dibahas pada bagian 2.1 di atas.

Setiap deskriptor yang dipakai mempunyai beberapa komponen. Komponen pertama adalah catatan cakupan (*scope note*), yang biasanya disingkat SN. Catatan cakupan ini bertujuan untuk menerangkan kepada para pemakai, mengenai arti maupun definisi singkat dari deskriptor tersebut.

Komponen kedua adalah adalah apa yang disebut dengan "digunakan untuk" (*use for*), biasanya disingkat dengan UF. Komponen "digunakan untuk" ini bertujuan untuk menginformasikan kepada para pemakai, bahwa deskriptor yang bersangkutan digunakan untuk mewakili istilah-istilah yang terdapat dalam komponen UF tersebut. Komponen ketiga, adalah komponen "gunakan" (*use*). Komponen ini merupakan kebalikan dari komponen kedua, yang bertujuan untuk menggiring para pemakai, untuk menggunakan standard deskriptor yang telah ditetapkan untuk

mewakilkan istilah-istilah tertentu. Komponen kedua dan ketiga ini bisa juga dipergunakan untuk mengatasi istilah-istilah yang sinonim.

Komponen keempat adalah "ungkapan lebih luas" (*broader terms*) dari suatu deskriptor. Biasanya "broader term" disingkat dengan BT. Tujuannya adalah untuk memberitahukan kepada para pemakai bahwa, deskriptor tersebut mempunyai istilah lain yang lebih generik dan lebih umum. Komponen yang kelima adalah "ungkapan lebih khusus" (*narrower terms*) dari suatu deskriptor, yang biasanya disingkat dengan NT. Ungkapan lebih khusus ini bertujuan untuk menginformasikan kepada para pemakai bahwa, deskriptor yang dipilih tersebut mempunyai istilah yang lebih spesifik. Komponen keempat dan kelima ini merupakan hubungan yang bersifat hirarkis. Suatu deskriptor bisa jadi mempunyai hubungan dengan deskriptor-deskriptor lainnya, dalam pengertian yang lebih luas, atau lebih khusus. Sebagai contohnya adalah hubungan yang bersifat "*whole/part*".

Komponen keenam adalah "ungkapan yang berhubungan" (*related terms*), dan biasanya disingkat dengan RT. Suatu deskriptor bisa jadi berhubungan dengan deskriptor lainnya, tanpa melalui hubungan yang bersifat hirarkis. Misalnya suatu deskriptor dapat berhubungan dengan deskriptor lainnya melalui fungsinya. Suatu deskriptor berikut semua komponennya dapat mewakili suatu konsep yang lebih lengkap. Sebagai contoh dapat dilihat pada gambar berikut ini:

Bilangan Rasional

SN : Digunakan untuk membilang bilangan bulat maupun pecahan
UF : Bilangan Nyata
Use:
BT : Sistem Bilangan
NT : Bilangan Integer
Bilangan Pecahan
Bilangan Logaritma
RT : Bilangan Imaginer

Gambar 1. Komponen dari suatu deskriptor

Tahap terakhir adalah menyusunan deskriptor-deskriptor yang menjadi entri dari suatu tesaurus menurut abjad.

2.3 Evaluasi Tesaurus

Setelah suatu tesaurus selesai dibuat, langkah selanjutnya adalah mengevaluasi tersaurus tersebut. Kembali kita terbentur, untuk mencari mekanisme yang baik untuk mengevaluasi efisiensi maupun efektivitas suatu tesaurus. Kalau sekiranya tesaurus tersebut merupakan suatu tesaurus yang mempunyai domain khusus, dan di sarikan dari sekelompok dokumen, maka cara untuk mengevalusi efektivitasnya adalah dengan memberikan beberapa pertanyaan (*query*), dan pertanyaan ini diterjemahkan ke dalam bahasa tesaurus (memakai deskriptor yang tersedia), kemudian dilihat apakah pertanyaan tersebut bisa terjawab dengan menemukan dokumen yang berkaitan dengan pertanyaan tersebut.

Cara yang lain adalah mendiskusikan tesaurus yang telah terbentuk tersebut kepada para pakar dalam bidang disiplin ilmu yang bersangkutan. Dalam inilah saya kira para pakar ahli bahasa banyak berperan, bukan saja untuk mengevaluasi, tapi juga dalam pemilihan kosakata yang tepat untuk dijadikan deskriptor.

Dalam proses evaluasi ini, perlu diantisipasi peng-update-an dari tesaurus tersebut. Tesaurus merupakan sesuatu alat yang terus berkembang untuk menyesuaikan diri dengan perkembangan informasi dan ilmu pengetahuan.

3. Peran Komputer Dalam Pembuatan Tesaurus

Pembuatan tesaurus memang tidak bisa sepenuhnya diserahkan kepada komputer. Seperti telah diuraikan di atas, ada beberapa aktivitas dalam pembuatan tesaurus yang tidak bisa didelegasikan kepada komputer, misalnya menentukan "*scope note*" dari suatu deskriptor, dan hubungan-hubungannya. Bagaimanapun juga, interfensi manusia (para pakar) diperlukan dalam hal ini.

Di samping itu ada beberapa aktivitas dalam pembuatan tesaurus, dengan menggunakan komputer, akan mempermudah pekerjaan si pembuat tesaurus. Aktivitas-aktivitas tersebut adalah:

1. Komputer bisa mempermudah penyusunan penampilan dari setiap entri yang digunakan.
2. Komputer bisa mempercepat penyortiran entri.
3. Komputer bisa mengkontrol hubungan hirarkis yang terdapat pada suatu entri.
4. Komputer dapat membantu dalam meng-update suatu tesaurus dengan fasilitas "*deletion*" dan "*insertion*".

Seperti telah diuraikan sebelumnya, hal yang perlu diperhatikan adalah pemilihan kosakata yang layak untuk dijadikan entri dalam suatu tesaurus. Dalam hal ini, komputer juga bisa membantu manusia secara otomatis, dalam menentukan kosakata apa yang tepat untuk dimuat di dalam tesaurus tersebut. Disamping komputer juga membantu manusia secara otomatis dalam menentukan kelas-kelas sub-disiplin yang mungkin terdapat pada sekumpulan dokumen dari suatu disiplin ilmu.

Riset mengenai penggunaan komputer dalam merancang suatu tesaurus sudah cukup banyak dilakukan orang. Diantara mereka yang pernah melakukan percobaan mengenai pembuatan tesaurus secara otomatik adalah Salton (1967), Wang (1985), Crouch (1988, 1990, 1992), Fox (1988), dan lain-lainnya. Riset-riset tersebut menggunakan statistik sebagai alat bantu dalam pemilihan kosakata secara otomatik. Statistik yang umumnya digunakan adalah analisis frekuensi kata-kata (*terms frequency analysis*) dari sekelompok dokumen, analisis dokumen kluster (*document cluster analysis*), dan analisis kluster untuk kosakata (*terms cluster analysis*). Pembuatan tesaurus secara otomatis ini, selalu dikaitkan dengan penelitian dalam bidang "*information retrieval*". Hasil penelitian meyakinkan bahwa dengan tersedianya tesaurus secara elektronis dalam suatu sistem temu-kembali informasi (*information retrieval system*), akan meningkatkan kinerja dari sistem tersebut.

4. Penutup

Tulisan ini merupakan salah satu suatu langkah awal dalam rangka pengembangan penelitian dalam bidang pemrosesan teks bahasa Indonesia di Fasilkom Universitas Indonesia. Tidak dapat disangkal lagi bahwa tulisan ini masih jauh dari sempurna, maka segala bantuan berupa saran yang membangun dari semua pihak, sangat diharapkan.

Pada saat ini, kami lagi berusaha mengembangkan tesaurus kecil dalam bidang ilmu komputer, sebagai salah satu komponen pengembangan sistem temu-kembali informasi. Tesaurus tersebut dikembangkan dari sekumpulan skripsi mahasiswa S-1 bidang Ilmu Komputer, Universitas Indonesia. Di harapkan, dari usaha yang kecil ini, akan membantu kami untuk mempunyai keyakinan yang lebih besar, untuk membuat tesaurus yang lebih sempurna, dalam rangka memperkaya khasanah penggunaan bahasa Indonesia.

5. Daftar Pustaka

Salton, G., R. T. Dattola, and D. M. Murray. 1967. "An Experiment in Automatic Thesaurus Construction". *Information Storage and Retrieval. Scientific Report* No. IRS-13 to the National Science Foundation. Ithaca, New York.

Townley, H. M. 1980. *Thesaurus Making*. Andre Deutsch Limited. London, United Kingdom.

Aitchison, J. 1970. "The Thesaurofacet: a Multi-purpose Retrieval Language Tool". *Journal of Documentation*, 26(3): 187-203.val". JASIS, 36(1):15-27

Lancaster, F. W. 1986. Vocabulary Control for Information Retrieval. 2nd ed. Information Resources. Arlington, VA.

Fox, E., et al. 1988. "Building a Large Thesaurus for Information Retrieval". Proceedings of the 2nd Conference on Applied Natural Language Processing. Austin, TX.

Crouch, C. 1988. "A Cluster-based Approach to Thesaurus Construction". Proceedings of the 11th International Conference on Research and Development in Information Retrieval. Grenoble, France.

Salton, G. 1989. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley Company. New York, NY.

Crouch, C. 1990. "An Approach to the Automatic Construction of Global Thesauri". Information Processing and Management, 26(5):629-640.

Crouch, C., et al. 1992. "Experiments in Automatic Statistical Thesaurus Construction". Proceedings of the 15th International Conference on Research and Development in Information Retrieval. Denmark.