# PERANCANGAN SUBSISTEM PENGOLAHAN PERTANYAAN UNTUK SISTEM RINGKASAN MULTI DOKUMEN BAHASA INDONESIA

**Putu Manik Prihatini**
Jurusan Teknik Elektro, Politeknik Negeri Bali
Bukit Jimbaran, P.O. Box 1064 Tuban Badung – BALI
Phone:+62-361-701981, Fax:+62-361-701128 E-mail: manikprihatini@pnb.ac.id

**Abstrak:** Ringkasan multi dokumen merupakan teknik yang diusulkan untuk memecahkan permasalahan pertanyaan kompleks. Pengolahan pertanyaan merupakan bagian penting dalam sistem ringkasan multi dokumen, karena pertanyaan harus dianalisis untuk mengetahui kebutuhan informasi yang dinginkan oleh pengguna. Kebutuhan informasi diketahui berdasarkan makna dari pertanyaan tersebut. Arsitektur subsistem pengolahan pertanyaan bahasa Indonesia pada penelitian ini dibangun berbasis topik, karena pertanyaan dari pengguna dapat direpresentasikan sebagai topik. Arsitektur subsistem pengolahan pertanyaan dirancang terdiri dari enam proses yaitu *tokenization*, *stop words*, *stemming*, penentuan sinonim, penentuan topik, dan analisis semantis. Melalui contoh kasus yang diberikan, dari proses pengolahan pertanyaan, dihasilkan topik dan makna pertanyaan yang akan digunakan sebagai masukan bagi subsistem berikutnya dalam sistem ringkasan multi dokumen bahasa Indonesia.

**Kata Kunci:** ringkasan multi dokumen, pengolahan pertanyaan, bahasa Indonesia

## *DESIGNING OF QUESTIONS PROCESSING SUB-SYSTEM FOR INDONESIAN MULTI DOCUMENT SUMMARIZATION SYSTEM*

***Abstract:*** *Multi document summarization is a proposed technique to solve the problem of complex questions. Questions processing is an important part in the multi document summarization, because the questions must be analyzed to determine the user's information needs. The information requirement is based on the meaning of questions. The architecture of Indonesian questions processing sub-system in this research is designed based on the topics, because the user's question can be represented as a topic. The architecture of Indonesian questions processing sub system consists of six processes, such as, tokenization, stop words, stemming, synonym determination, topics determination, and semantic analysis. Through the case given, by questions processing, resulted topics and meaning of the questions that will be used as input for the next sub-system in Indonesian multi document summarization.*

***Key words****: multi document summarization, questions processing, Indonesian*

## I. PENDAHULUAN

Manusia sebagai makluk pribadi dan sosial selalu membutuhkan informasi, baik untuk kepentingan dirinya sendiri maupun untuk menjalin hubungan dengan makluk sosial lainnya. Semakin canggihnya perkembangan ilmu pengetahuan dan teknologi telah menghadirkan fenomena baru dalam penyediaan informasi, tidak hanya melalui media cetak, tetapi tersedia juga secara *online*, bahkan dalam jumlah yang sangat banyak. Perkembangan ilmu pengetahuan dan teknologi di bidang pencarian informasi telah menunjukkan kemampuan komputer sebagai mesin yang dapat mempermudah manusia mendapatkan informasi melalui sumber-sumber yang tersedia secara *online*. Akan tetapi, sistem pencarian yang selama ini tersedia belum mampu memenuhi kebutuhan pengguna sistem terhadap hasil pencarian yang didapatkan [6]. Keterbatasan dalam teknologi mesin pencarian informasi ini membutuhkan suatu mekanisme yang memudahkan pengguna memperoleh informasi yang merupakan gabungan dari informasi-informasi yang saling berhubungan dari beberapa dokumen pada *website* yang berbeda.

Perkembangan ilmu pengetahuan di bidang pencarian informasi pada *web* menghasilkan suatu tugas khusus untuk menggabungkan informasi dari beberapa dokumen berbeda yang disebut dengan Ringkasan Multi Dokumen [13]. Ringkasan multi dokumen menghasilkan suatu ringkasan dari beberapa dokumen dengan topik yang sama [12]. Ringkasan berbasis topik merupakan bentuk khusus dari ringkasan multi dokumen, ringkasan dibuat dari sekumpulan dokumen untuk menjawab kebutuhan informasi yang dinyatakan dalam bentuk pertanyaan dari pengguna [17, 21]. Pertanyaan yang dimaksud di sini adalah kriteria yang diketikkan oleh pengguna pada mesin pencarian informasi.

Pertanyaan dari pengguna perlu diolah melalui mekanisme khusus untuk menentukan kebutuhan informasi yang diinginkan oleh pengguna. Pengolahan pertanyaan menangani beberapa tipe pertanyaan seperti ya-tidak, fakta, definisi, daftar, bagaimana, mengapa, hipotesis, lintas bahasa, instruksi dan penjelasan [1, 5, 11, 14]. Beberapa tipe pertanyaan merupakan pertanyaan sederhana, karena jawabannya dapat diperoleh dari sebagian kecil teks dokumen, sedangkan pertanyaan kompleks membutuhkan kombinasi dari beberapa tipe informasi dan bukan berarti bahwa suatu jawaban tunggal dapat memenuhi semua kebutuhan informasi [4-7]. Ringkasan multi dokumen merupakan teknik yang diusulkan untuk memecahkan permasalahan pertanyaan kompleks [2, 3, 5, 8, 10].

Penelitian-penelitian telah dilakukan untuk mengolah pertanyaan dalam bahasa Indonesia menggunakan beberapa metode yaitu *deeper linguistic approach* [15], *machine learning approach* [18], *linguistic and world knowledge axioms* [16], *monolingual approach* [22], dan *phrase-based approach* [23]. Dari penelitian-penelitian tersebut, belum ada penelitian yang mengolah pertanyaan untuk ringkasan multi dokumen sebagai solusi dalam menjawab pertanyaan kompleks. Untuk itu, melalui penelitian ini, penulis ingin merancang arsitektur subsistem pengolahan pertanyaan sebagai bagian penting dari sistem ringkasan multi dokumen bahasa Indonesia.

## II. PENCARIAN INFORMASI

Pencarian informasi dapat diartikan sebagai aplikasi dari teknologi komputer untuk memperoleh, mengorganisir, menyimpan, mencari dan menyebarkan informasi [13]. Pencarian informasi merupakan tugas untuk menemukan dokumen yang sesuai dengan kebutuhan pengguna [20].

Karakteristik dari sistem pencarian informasi adalah:

1. Pengumpulan dokumen, setiap sistem harus memutuskan apa yang akan digunakan sebagai dokumen, apakah suatu paragraf, suatu halaman, atau beberapa halaman,
2. Pertanyaan dalam bahasa *query*, menyatakan apa yang dibutuhkan pengguna, dapat berupa daftar kata-kata, suatu frasa dari kata-kata, operator Boolean, atau bukan operator Boolean,
3. Sekumpulan hasil, merupakan bagian dari dokumen yang dianggap oleh sistem pencarian informasi sesuai dengan pertanyaan,
4. Penyajian terhadap sekumpulan hasil, secara sederhana dinyatakan dalam bentuk perangkingan daftar judul dokumen.

## III. RINGKASAN MULTI DOKUMEN

Ringkasan multi dokumen bertujuan untuk menghasilkan presentasi yang berasal dari beberapa dokumen, atau dari sisi teknik pembuatan ringkasan, bertujuan untuk mengambil fitur kunci dan fondasi dasar dari suatu bidang tertentu, perkembangan awal dan terakhir, kontribusi dan penemuan penting, posisi kontradiksi yang mungkin mengembalikan tren atau memulai sub bidang baru, dan definisi dasar serta contoh yang dapat dipahami dengan cepat oleh pembaca yang bukan ahlinya [19]. Ringkasan multi dokumen menghasilkan ringkasan informasi dari sekumpulan dokumen yang memiliki topik utama baik implisit maupun eksplisit, informasi terdapat dalam suatu kluster dokumen dan membantu pengguna memahami kluster dokumen [21]. Idealnya, ringkasan multi dokumen berisi satu informasi kunci relevan yang sama dari semua dokumen, ditambah dengan informasi unik dari beberapa dokumen yang sesuai dengan kebutuhan pengguna [9].

Sistem ringkasan dokumen otomatis dikembangkan dalam tiga tahapan sebagai berikut [12]. Tahap pertama adalah identifikasi topik, menghasilkan tipe ringkasan paling sederhana. Topik adalah subyek tertentu yang ditulis atau didiskusikan. Tahap kedua adalah interpretasi, sintesis terhadap konsep, evaluasi dan proses lainnya. Hasil dari interpretasi biasanya berupa representasi abstrak yang tidak bisa dibaca pengguna, dan bahkan hasil ekstraksinya kadang koheren, disebabkan oleh untaian referensi, penghapusan sebagian hubungan wacana, dan pengulangan atau penghapusan sebagian materi. Tahap ketiga adalah pembangkitan ringkasan, menghasilkan teks yang bisa dibaca oleh pengguna.

## IV. HASIL DAN PEMBAHASAN

Pengolahan pertanyaan merupakan bagian penting dalam sistem ringkasan multi dokumen, karena pertanyaan harus dianalisis untuk mengetahui kebutuhan informasi yang dinginkan oleh pengguna. Kebutuhan informasi diketahui berdasarkan makna dari pertanyaan tersebut. Untuk itu, pengolahan pertanyaan membutuhkan mekanisme khusus agar dapat menganalisis pertanyaan dari pengguna. Pada penelitian ini, pengolahan pertanyaan dirancang sebagai sebuah subsistem dari sistem ringkasan multi dokumen secara keseluruhan.

Arsitektur subsistem pengolahan pertanyaan bahasa Indonesia ini dibangun berbasis topik, karena pertanyaan dari pengguna dapat direpresentasikan sebagai topik. Selanjutnya, topik digunakan sebagai dasar dalam pencarian dokumen-dokumen yang sesuai. Informasi dalam

dokumen-dokumen diekstrak untuk menghasilkan sebuah ringkasan sebagai jawaban atas pertanyaan pengguna. Arsitektur subsistem pengolahan pertanyaan dapat dilihat pada Gambar 1.

Arsitektur subsistem pengolahan pertanyaan bahasa Indonesia dirancang terdiri dari enam proses yaitu (1) *tokenization*, (2) *stop words*, (3) *stemming*, (4) penentuan sinonim, (5) penentuan topik, dan (6) analisis semantis.

Sebagai contoh, diberikan teks pertanyaan sebagai berikut.

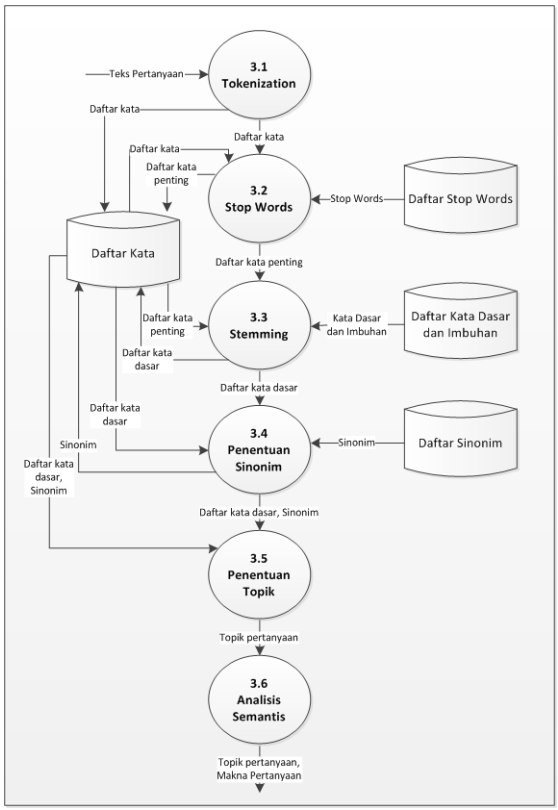
| **Saya ingin mencari informasi tentang gejala kanker rahim** |
|---|

### 4.1 *Tokenization*

*Tokenization* merupakan proses untuk menguraikan teks dokumen menjadi kalimat dan menguraikan kalimat menjadi kata-kata. Untuk contoh kasus di atas, proses ini menghasilkan daftar kata-kata sebagai berikut.

| **saya** | **ingin** | **mencari** | **informasi** | **tentang** |
|---|---|---|---|---|
| **gejala** | **kanker** | **rahim** | | |

Seluruh kata-kata di atas akan disimpan dalam basis data daftar kata sebagai daftar kata.




Gambar 1. Arsitektur Subsistem Pengolahan Pertanyaan

### 4.2 *Stop Words*

*Stop words* merupakan proses untuk menghilangkan kata-kata yang tidak penting seperti spasi, kata hubung, kata depan, dan lain-lain. Daftar *stop words* telah tersimpan dalam basis data Daftar *Stop Words*. Untuk contoh kasus di atas, proses ini melakukan penghapusan terhadap kata "saya", "ingin", dan "tentang", sehingga menghasilkan daftar kata-kata sebagai berikut.

| **Mencari** | **informasi** | **gejala** | **kanker** | **rahim** |
|---|---|---|---|---|

Seluruh kata-kata di atas akan disimpan kembali ke dalam basis data daftar kata sebagai daftar kata penting.

### 4.3 *Stemming*

*Stemming* merupakan proses untuk menguraikan sebuah kata menjadi kata dasarnya. Daftar kata dasar dan imbuhan telah tersimpan dalam basis data daftar kata dasar dan imbuhan. Untuk contoh kasus di atas, proses ini menguraikan kata "mencari" menjadi kata dasarnya yaitu "cari", sehingga menghasilkan daftar kata sebagai berikut.

| **cari** | **informasi** | **gejala** | **kanker** | **rahim** |
|---|---|---|---|---|

Seluruh kata dasar di atas akan disimpan kembali ke dalam basis data daftar kata sebagai daftar kata dasar.

### 4.4 Penentuan Sinonim

Penentuan sinonim merupakan proses untuk mencari padanan kata dari suatu kata. Penentuan sinonim dilakukan untuk mengatasi perbedaan kata pada dokumen sumber dengan kata pada teks pertanyaan, padahal memiliki makna yang sama. Penentuan sinonim dilakukan terhadap kata dasar maupun turunannya. Daftar sinonim telah tersimpan dalam basis data daftar sinonim.

| Kata | Sinonim |
|---|---|
| Cari | gagar, geledah, kerekau, kerosek, mencekau, raba, |
| Mencari | berburu, melacak, memancing, membongkar-bongkar, memecahkan, memeriksa, memilih, memisah-misahkan, menangkap, menduga, mengacar, mengaduk-aduk, mengagau, mengebek, mengejar, mengetahui, menggali, menggeledah, menggeratak, mengungkai, mengusut, menyelam, menyelesaikan, menyelidik, menyidik, menyondong, meraba, meranya, merayau, merisik, |
| Informasi | penerangan, penjelasan, bahan, berita, data, embaran, fakta, kabar, keterangan, laporan, liputan, warta |
| Gejala | fakta, fenomena, kenyataan, alamat, gelagat, indikasi, isyarat, naga-naga, pelebaya, pertanda, petunjuk, riak, simtom, sinyal, tanda-tanda |
| Kanker | puru ajal, tumor ganas |
| Rahim | kandungan, peranakan |

Untuk contoh kasus di atas, proses ini menghasilkan sinonim terhadap seluruh kata dasar dan turunannya sebagai berikut. Seluruh sinonim kata di atas akan disimpan kembali ke dalam basis data daftar kata sebagai sinonim.

### 4.5 Penentuan Topik

Penentuan topik merupakan proses untuk menganalisis maksud dari pertanyaan. Salah satunya adalah mencari peran semantis dari kata-kata dalam teks pertanyaan. Untuk contoh di atas, kata "cari" dan "informasi" mempunyai peran semantis keterangan terhadap gabungan kata "gejala", "kanker" dan "rahim". Untuk itu, kata "cari" dan "informasi" dapat dihilangkan, sehingga topik dari teks pertanyaan pada contoh di atas adalah seperti berikut.

| Kata | Sinonim |
|---|---|
| gejala | fakta, fenomena, kenyataan, alamat, gelagat, indikasi, isyarat, naga-naga, pelebaya, pertanda, petunjuk, riak, simtom, sinyal, tanda-tanda |
| kanker | puru ajal, tumor ganas |
| rahim | kandungan, peranakan |

Seluruh daftar kata dasar dan sinonim kata di atas akan disimpan kembali ke dalam basis data daftar kata sebagai daftar kata dasar dan sinonim.

### 4.6 Analisis Semantis

Analisis semantis merupakan proses penggabungan kembali daftar kata menjadi frasa kata atau kalimat untuk dianalisis hubungan semantisnya, sehingga menghasilkan makna dari pertanyaan berdasarkan topik pertanyaan. Untuk contoh kasus di atas, seluruh kata dasar akan digabungkan menjadi frasa kata benda sebagai berikut.

"**gejala kanker rahim**"

Proses ini selanjutnya menganalisis hubungan semantis dari frasa tersebut. Salah satu contoh hasil hubungan semantisnya adalah sebagai berikut.

"**gejala yang dialami jika seseorang mengalami penyakit kanker rahim**"

Jadi, dari proses pengolahan pertanyaan berdasarkan contoh kasus diatas, dihasilkan topik dan makna pertanyaan yang akan digunakan sebagai masukan bagi subsistem berikutnya dalam sistem ringkasan multi dokumen bahasa Indonesia.

## V. SIMPULAN DAN SARAN

### 5.1 Simpulan

Dari pembahasan yang dilakukan dapat diperoleh kesimpulan bahwa subsistem pengolahan pertanyaan memegang peranan penting dalam sistem ringkasan karena pertanyaan harus dianalisis untuk mengetahui kebutuhan informasi yang dinginkan oleh pengguna. Kebutuhan informasi diketahui berdasarkan makna dari pertanyaan tersebut. Arsitektur subsistem pengolahan pertanyaan bahasa Indonesia pada penelitian ini dibangun berbasis topik, karena pertanyaan dari pengguna dapat direpresentasikan sebagai topik. Arsitektur subsistem pengolahan pertanyaan dirancang terdiri dari enam proses yaitu *tokenization*, *stop words*, *stemming*, penentuan sinonim, penentuan topik, dan analisis semantis. Melalui contoh kasus yang diberikan, dari proses pengolahan pertanyaan, dihasilkan topik dan makna pertanyaan yang akan digunakan sebagai masukan bagi subsistem berikutnya dalam sistem ringkasan multi dokumen bahasa Indonesia.

### 5.2 Saran

Untuk mengetahui unjuk kerja dari rancangan arsitektur subsistem pengolahan pertanyaan bahasa Indonesia dalam penelitian ini, disarankan untuk mengimplementasikan rancangan menjadi subsistem berbasis komputer sehingga dapat diukur tingkat ketepatan topik dan makna dari pertanyaan yang dihasilkan.

## DAFTAR PUSTAKA

[1] Bdour, W.N. and N.K. Gharaibeh, *Development of Yes/No Arabic Question Answering System.* International Journal of Artificial Intelligence & Applications, 2013. **4**(1): p. 51-63.

[2] Chali, Y. and S.A. Hasan, *Query focused multidocument summarization: automatic data annotations and supervised learning approaches.* Natural Language Engineering, 2012. **18**(1): p. 109-145.

[3] Chali, Y., S.A. Hasan, and K. Imam. *Improving the Performance of the Reinforcement Learning Model for Answering Complex Questions.* in *Proceedings of the 21st ACM conference on information and knowledge management (CIKM 2012).* 2012. Maui, HI, USA: ACM.

[4] Chali, Y., S.A. Hasan, and S.R. Joty. *Do Automatic Annotation Techniques Have Any Impact on Supervised Complex Question Answering?* in *Proceedings of the joint conference of the 47th annual*

*meeting of the association for computational linguistics*. 2009. Singapore: Suntec.

[5] Chali, Y., S.A. Hasan, and S.R. Joty, *Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels.* Information Processing and Management, 2011. **47**(6): p. 843-855.

[6] Chali, Y., S.A. Hasan, and M. Mojahid, *A reinforcement learning formulation to the complex question answering problem.* Information Processing and Management, 2015. **51**: p. 252-272.

[7] Chali, Y., S.R. Joty, and S.A. Hasan, *Complex Question Answering: Unsupervised Learning Approaches and Experiments.* Journal of Artificial Intelligence Research, 2009. **35**: p. 1-47.

[8] Figueroa, A., G. Neumann, and J. Atkinson, *Searching for Definitional Answers on the Web Using Surface Patterns*, in *IEEE Computer Society*2009. p. 68-76.

[9] Goldstein, J., et al. *Multi-Document Summarization By Sentence Extraction*. in *NAACL-ANLP-AutoSum '00 Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization*. 2000. Stroudsburg, PA, USA: Association for Computational Linguistics ACM.

[10] Harabagiu, S., F. Lacatusu, and A. Hickl. *Answering Complex Questions with Random Walk Models*. in *SIGIR '06*. 2006. Seattle, Washington, USA: ACM.

[11] He, Z. and E. Lo, *Answering Why-Not Questions on Top-K Queries.* IEEE Transactions on Knowledge and Data Engineering, 2014. **26**(6): p. 1300-1315.

[12] Hovy, E., *Text Summarization*, in *The Oxford Handbook of Computational Linguistics* R. Mitkov, Editor 2005, OUP Oxford. p. 583-598.

[13] Jackson, P. and I. Moulinier, *Natural Language Processing for Online Applications : text retrieval, extraction, and categorization*2002, Amsterdam: John Benjamins B.V.

[14] Khillare, S.A., B.A. Shelke, and N. Mahender, *Comparative Study on Question Answering Systems and Techniques.* International Journal of Advanced Research in Computer Science and Software Engineering, 2014. **4**(11): p. 775-778.

[15] Larasati, S.D. and R. Manurung. *Towards a Semantic Analysis of Bahasa Indonesia for Question Answering*. in *Proceedings of the 10th Conference of the Pasific Association for Computational Linguistics (PACLING 2007)*. 2007.

[16] Mahendra, R., S.D. Larasati, and R. Manurung. *Extending an Indonesian Semantic Analysis-based Question Answering System with Linguistic and World Knowledge Axioms*. in *22nd Pacific Asia Conference on Language, Information and Computation*. 2008.

[17] Nastase, V. *Topic-Driven Multi-Document Summarization with Encyclopedic Knowledge and Spreading Activation*. in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 2008. Honolulu: Association for Computational Linguistics.

[18] Purwarianti, A., M. Tsuchiya, and S. Nakagawa. *A Machine Learning Approach for Indonesian Question Answering System*. in *Proceedings of the International Conference on Artificial Intelligence and Applications (AIA 2007)*. 2007.

[19] Qazvinian, V., et al., *Generating Extractive Summaries of Scientic Paradigms.* Journal of Articial Intelligence Research, 2013. **46**(165-201).

[20] Russell, S. and P. Norvig, *Artificial Intelligence A Modern Approach*. Second Edition2003, New Jersey: Pearson Education, Inc. 790.

[21] Wan, X., J. Yang, and J. Xiao. *Manifold-Ranking Based Topic-Focused Multi-Document Summarization*. in *IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence* 2007. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ACM.

[22] Zulen, A.A. and A. Purwarianti. *Study and Implementation of Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question*. in *25th Pacific Asia Conference on Language, Information and Computation*. 2011.

[23] Zulen, A.A. and A. Purwarianti. *Using Phrase-Based Approach in Machine Learning Based Factoid Indonesian Question Answering*. in *CISAK2013*. 2013.