PREDIKSI TOPIK PENELITIAN MENGGUNAKAN KOMBINASI ANTARA SUPPORT VECTOR REGRESSION DAN KURVA LOGISTIK

Agus Widodo, Indra Budi, Rizal Fathoni Aji
Fakultas Ilmu Komputer, Universitas Indonesia
Kampus UI Depok 16424
Telp. (021) 786 3419, Faks. (021) 786 34152
E-mail: agus.widodo@ui.ac.id, indra@cs.ui.ac.id, rizal@cs.ui.ac.id

ABSTRAK
Prediksi topik penelitian diperlukan terutama oleh pengambil kebijakan misalnya untuk menentukan tema penelitian yang akan datang. Analisis terhadap tren topik penelitian dengan menggunakan pendekatan numerik berdasarkan publikasi ilmiah dan/atau paten telah dilakukan dalam beberapa peneliti sebelumnya dengan menghitung frekuensi kata yang sama dalam beberapa dokumen untuk mengelompokkan topik. Sementara itu, analisis time series dilakukan peneliti lainnya dengan menggunakan teknik statistik, pembelajaran mesin dan kurva logistik untuk domain di luar topik penelitian. Dalam penelitian ini, akan dilakukan analisis time series untuk memprediksi topik penelitian dengan menggabungkan teknik pembelajaran mesin, yakni Support Vector Regression (SVR), dan kurva logistik, dengan asumsi bahwa terdapat batas atas terhadap siklus pertumbuhan setiap topik penelitian. Data yang digunakan dalam studi ini adalah data laporan penelitian dari situs Garuda selama kurun waktu 16 tahun. Selain itu, sebagai acuan, digunakan juga data dari ITU (International Telecommunication Unit) yang serupa dengan yang digunakan oleh peneliti sebelumnya. Hasil eksperimen menunjukkan bahwa kombinasi antara SVR dan kurva logistik dapat meningkatkan akurasi prediksi. Selain itu, dari data yang digunakan, dapat diketahui bahwa topik penelitian yang pertumbuhannya cukup tinggi adalah Medicine, Biology *dan* Social pathology and public walfare.

Kata kunci: kombinasi prediksi, topik penelitian, kurva logistik, support vector regression

1. PENDAHULUAN

Peneliti maupun pengambil kebijakan perlu memahami kondisi riset saat ini dan pada masa yang akan datang, serta mampu mengidentifikasi area riset yang memiliki potensi besar. Sementara itu, sumber-sumber informasi ilmu pengetahuan dan teknologi (*Science and Technologi/S&T*) saat ini telah berkembang dengan pesat seiring dengan kemajuan internet. Sumber informasi tentang S&T yang utama meliputi basis data tentang publikasi ilmiah dan paten.

Berdasarkan jumlah paten dan publikasi ilmiah tersebut, teknologi dapat dikategorikan menjadi teknologi yang sedang tumbuh (*emerging technology*), teknologi yang relatif stabil, dan teknologi yang sudah tidak ada atau sedikit sekali aktivitas (Bengisu dan Nekhili, 2006). Teknologi yang sedang tumbuh akan menjadi teknologi kunci bagi industri pada masa yang akan datang.

Sementara itu, beberapa cara untuk melakukan prediksi, secara garis besar bisa dikategorikan menjadi analisis *judgemental* dan kuantitatif (Meredith, 1995). Prediksi berbasis data numerik mengekstrapolasikan data historis melalui suatu fungsi tertentu, sedangkan peramalan *judgemental* bisa juga didasarkan pada proyeksi dari masa lalu, tetapi sumber-sumber informasi dalam model tersebut bergantung pada penilaian subjektif para ahli. Dalam (Bengisu dan Nekhili, 2006) dinyatakan bahwa hasil panel para ahli untuk studi peramalan melalui analisis Delphi sebagian tidak bersesuaian dengan hasil analisis numerik. Karena keterwakilan ahli dalam panel yang kurang proporsional bisa mengurangi akurasi prediksi.

Analisis terhadap kecenderungan topik penelitian dengan menggunakan pendekatan numerik berdasarkan publikasi ilmiah dan/atau paten telah dilakukan dalam beberapa peneliti sebelumnya, misalnya Small (2006), Rahayu dan Hasibuan (2006), and Daim (2006), Woon, Hensche dan Madnick (2009), Ziegler (2009), serta Vidican et al. (2009). Peneliti-peneliti tersebut menghitung kata yang sama dalam beberapa dokumen untuk mengelompokkan teknologi, dan menghitung frekuensi kata untuk menentukan tingkat kecenderungan teknologi.

Analisis *time series* terhadap publikasi ilmiah dan/atau paten dilakukan oleh beberapa peneliti dengan teknik yang beragam. Misalnya, Bengisu dan Nekhili (2006) dengan menggunakan kurva logistik, sedangkan Jun dan Uhm (2010) menggunakan pendekatan statistik dan teknik pembelajaran mesin terhadap data paten. Pada penelitian sebelumnya, penulis (Widodo, Fanany dan Budi, 2011) menggunakan analisis time series untuk melakukan prediksi terhadap data penelitian dari PubMed, dan teknik pembelajaran mesin dapat memberikan kinerja yang lebih baik dari pendekatan statistik. Sementara itu, penggabungan antara kurva logistik dan teknik statistik (ARIMA) dilakukan oleh Christodulous et al. (2010) untuk melakukan prediksi dengan data yang terbatas.

Pada penelitian ini, akan dilakukan analisis *time series* untuk memprediksi topik penelitian dengan menggabungkan teknik pembelajaran mesin, yakni *Support Vector Regression* (SVR) dan kurva logistik. Penggunaan teknik pembelajaran mesin didasarkan pada kelebihan teknik ini dalam prediksi data time series, sedangkan penggunaan kurva logistik didasarkan pada asumsi bahwa topik penelitian yang memiliki batasan dalam siklus pertumbuhannya.

Selanjutnya, sistematika penulisan diawali dengan review literatur terhadap penelitian terkait sebelumnya, kemudian dilanjutkan dengan teori tentang SVR dan kurva logistik. Sub-bab selanjutnya akan membahas metodologi yang digunakan, yang meliputi penentuan setting untuk pengukuran, termasuk penjelasan data, algoritma yang digunakan, dan pengukuran kinerja. Sedangkan pada sub-bab hasil dan pembahasan akan disampaikan hasil uji coba yang dilakukan.

2. REVIEW LITERATUR

2.1 Prediksi Topik Penelitian

Kajian tentang topik penelitian telah dilakukan oleh beberapa peneliti sebelumnya. Small (2006) menggunakan *co-citation* untuk melakukan pengelompokan area riset di bidang *science*, sedangkan Rahayu dan Hasibuan (2006) dan Zhu et al. (1998) menggunakan *co-word* analysis. Untuk menentukan kategori topik penelitian yang sedang tumbuh, Rahayu dan Hasibuan (2006) menggunakan batas persentase tertentu, yakni antara satu sampai tiga persen, terhadap total jumlah paten atau publikasi ilmilah.

Woon et al. (2009) melakukan penelitian tentang tren teknologi menggunakan *bibliometrics*, yakni jumlah record yang dihasilkan dari suatu query terhadap basis data online. Dari judul dan abstrak yang diperoleh, diperoleh kata kunci (*terms*) yang kemudian disusun dalam taksonomi dengan menggunakan *Normalized Google Distance*. Skor pertumbuhan teknologi dihitung berdasarkan rata-rata tahun publikasi (*average publication year*), yakni frekuensi istilah dikalikan dengan tahun dibagi dengan frekuensi istilah selama periode observasi. Dengan demikian, tahun terakhir akan memberikan dampak yang lebih besar pada ukuran pertumbuhan.

Sebelumnya, Kobayashi et al. (2005) melakukan prediksi terhadap teknologi tertentu menggunakan informasi yang tersedia secara online. Istilah dalam kamus digunakan untuk menyaring kata kunci yang didapatkan dari dokumen hasil pencarian. Dalam tulisan ini, digunakan teknik untuk menggabungkan keywords yang serupa, teknik untuk menentukan hirarki antar keywords dan teknik untuk menentukan tingkat kepentingan suatu keywords.

2.2 Penggunaan Kurva Logistik

Perdapat beberapa penelitian yang menggunakan kurva logistik untuk memprediksi tren teknologi masa depan. Christodoulos et al. (2010) menggabungkan metode statistik, yakni ARIMA (*Autoregressive Integrated Moving Average*) dengan kurva logistik untuk memprediksi tingkat difusi teknologi. Dua pendekatan ini dikembangkan dalam perspektif riset dan fenomena yang berbeda. Model difusi berasal dari ilmu biologi, ekonomi industri, dan ekonomi bisnis. Di lain pihak, model ARIMA berasal dari ilmu matematika dan statistik, yang biasanya digunakan untuk aplikasi prediksi setelah didapatkan sejumlah data yang cukup banyak.

Penelitian tersebut mengindikasikan bahwa metode ARIMA lebih baik untuk prediksi jangka pendek, sedangkan kurva logistik lebih baik untuk prediksi jangka panjang. Hasil prediksi ARIMA kemudian digunakan sebagai data aktual untuk membangun model difusi. Penulis ini berkesimpulan bahwa kombinasi dari dua pendekatan tersebut menghasilkan kinerja yang lebih baik daripada kinerja masing-masing prediktor. Penulis menyatakan bahwa kekurangan dari metodologi yang digunakan adalah persyaratan data historis yang cukup untuk membentuk time series, dan jangka waktu prediksi masih satu tahun ke depan.

Daim et al. (2006) melakukan prediksi dalam tiga bidang teknologi dengan mengintegrasikan penggunaan bibliometrics dan analisis paten ke dalam teknik peramalan teknologi (*technology forecasting*), yakni perencanaan skenario, kurva pertumbuhan (*S-curve*) dan analogi. Sumber data dikelompokkan menjadi beberapa kategori jenis litbang, misalnya *Science Citation Index* untuk riset dasar, *Engineering Index* untuk riset terapan, *US Patent* untuk tahap pengembangan, *Newspaper Abstract Daily* untuk tahap aplikasi dan *Business and Popular Press* untuk tahap dampak sosial.

Bengisu dan Nekhili (2006) melakukan prediksi 20 jenis teknologi dalam kategori "*machine and materials*" yang sedang tumbuh dengan menggunakan jumlah paten dan publikasi selama 11 tahun (1994 - 2004). Proyeksi jangka pendek terhadap beberapa teknologi terpilih dilakukan menggunakan kurva pertumbuhan Gompertz atau logistic. Hasil penelitian menunjukkan bahwa sebagian besar teknologi yang diprediksi oleh pakar dengan metode Delphi memang dalam tahap sedang tumbuh, tetapi sebagian lainnya sebenarnya dalam tahap yang relative *idle*.

3. LANDASAN TEORI

3.1 Support Vector Regression

Support Vector Regression (SVR) yang merupakan Support Vector Machines (SVM) untuk regresi yang merepresentasikan fungsi dengan sebagian training data, yang biasa disebut support vectors (Phienthrakul et al., 2006). Muller et al. (1997) menyatakan bahwa SVM menunjukkan performance yang sangat baik untuk prediksi time series. Jika terdapat training data $\{(x_1, y_1), K, (x_1, y_1)\} \subset X \times R$, dimana X menyatakan pola input, SVM

mencari suatu fungsi *f(x)* yang memiliki deviasi maksimal ε dari nilai target y_i yang sebenarnya. Untuk suatu fungsi f yang linear bisa dituliskan dalam

$$f(x) = \langle w, x \rangle + b \text{ with } w \in X, b \in R \quad (1)$$

Jika beberapa kesalahan bisa ditoleransi, maka fungsi *error soft margin* biasa digunakan, dimana terdapat suatu variabel *slack* ξ yang memungkinkan fungsi optimasi bekerja dengan baik.

Teknik yang memungkinkan SVM menghasilkan prediksi *non-linear* adalah dengan melakukan pemetaan ruang input ke dalam ruang fitur berdimensi tinggi melalui suatu fungsi mapping Φ, dengan mengganti masing-masing data training x_i dengan $\Phi(x_i)$. Akan tetapi, bentuk eksplisit dari Φ tidak perlu diketahui, cukup dengan suatu inner product dalam ruang fitur, yang biasa disebut dengan fungsi kernel, $K(x,y) = \Phi(x) \cdot \Phi(y)$.

Beberapa kernel yang biasa digunakan adalah Gausian Radial Basis Function, Polynomial atau Linear. Masing-masing kernel berhubungan dengan suatu ruang fitur dan karena mapping eksplisit tidak perlu dilakukan, maka pemisah linear yang optimal dapat diperoleh secara efisien meskipun dalam suatu ruang fitur dengan dimensi yang sangat banyak (Cristianini, 2001).

3.2 Kuva Logistik

Metode kurva logistik, yang biasa juga disebut dengan kurva-S atau kurva pertumbuhan, menggunakan pendekatan bahwa pertumbuhan biasanya berbentuk lengkung seperti huruf S yang dapat dibagi menjadi tiga tahap. Yang pertama adalah pertumbuhan awal yang lambat, yang kemudian diikuti oleh masa pertumbuhan yang cepat dan pada akhirnya bergerak landai ke arah batas atas. Fungsi matematis atau model biasanya digunakan untuk ekstrapolasi ini.

Kurva logistik biasa digunakan dalam pemodelan pertumbuhan species dalam biologi dengan persamaan (Mathews, 1992)

$$f(t) = \frac{K}{1+e^{-r(t-t_0)}} \quad (2)$$

dimana K adalah batas atas atau *carrying capacity* dari teknologi tertentu, r adalah tingkat awal kecepatan pertumbuhan, dan t_0 adalah waktu awal pertumbuhan. Nilai K biasanya didapatkan dari data sekunder atau pendapat ahli. Jika nilai batas atas K tidak diketahui, maka bisa diestimasi dengan meminimalkan *sum of squared error* (Mathews, 1992)

$$E(r, K) = \sum_{k=1}^{n} \left[\frac{K}{1+e^{-r(t_k-t_0)}} - p_k\right]^2 \quad (3)$$

dimana p_k adalah nilai dari data yang diketahui.

3.3 Tingkat Pertumbuhan

Woon dan Ziegler (2009) menggunakan beberapa alternatif untuk menghitung tingkat pertumbuhan (*growth rate*) dari suatu topik penelitian, yakni (1) perbedaan antara frekuensi pada tahun terakhir dan tahun awal, (2) rasio antara frekuensi pada tahun terakhir dan tahun awal, (3) tingkat fitting terhadap kurva eksponensial, dan (4) rata-rata tahun publikasi. Untuk memberikan hasil yang lebih berimbang, maka frekuensi istilah tertentu dapat dinormalisasi dengan membagi frekuensi tersebut dengan jumlah total publikasi pada tahun tertentu.

Fitting terhadap kurva eksponensial akan menghasilkan bentuk $a \times e^r$, dimana r merupakan ukuran tingkat pertumbuhan. Sedangkan rata-rata tahun publikasi dihitung dengan menjumlah hasil perkalian antara tahun dan frekuensi pada tahun tersebut dibagi dengan total jumlah frekuensi. Sehingga,

$$avg\ year\ of\ pub = \sum_i (y_i \times h_i) / \sum_i h_i. \quad (4)$$

Publikasi tahun terakhir akan memiliki bobot yang lebih tinggi dari tahun-tahun sebelumnya.

3.4 Kombinasi Prediksi

Beberapa teknik kombinasi yang biasa dilakukan adalah melalui rata-rata, baik sederhana maupun bobot berdasarkan kinerja *predictor*, dan pemeringkatan.

Dalam teknik rata-rata, hasil prediksi akhir dihitung berdasarkan rata-rata dari hasil prediksi individu. Cara yang paling sederhana adalah dengan bobot yang sama, dimana hasil prediksi dari vektor *x* dari sebanyak *M* prediktor adalah

$$\bar{x} = \frac{1}{M} \sum_1^M x_i \quad (5)$$

Proses rata-rata ini bisa mengurangi tingkat kesalahan dalam prediksi jika semua prediktor memiliki akurasi yang setara. Jika tidak, sebaiknya digunakan rata-rata dengan bobot tertimbang, dimana bobot dihitung berdasarkan kinerja dari proses training atau validasi,

$$w_i = \frac{1/MSE_i}{\sum_{i=1}^{M} 1/MSE_i} \quad (6)$$

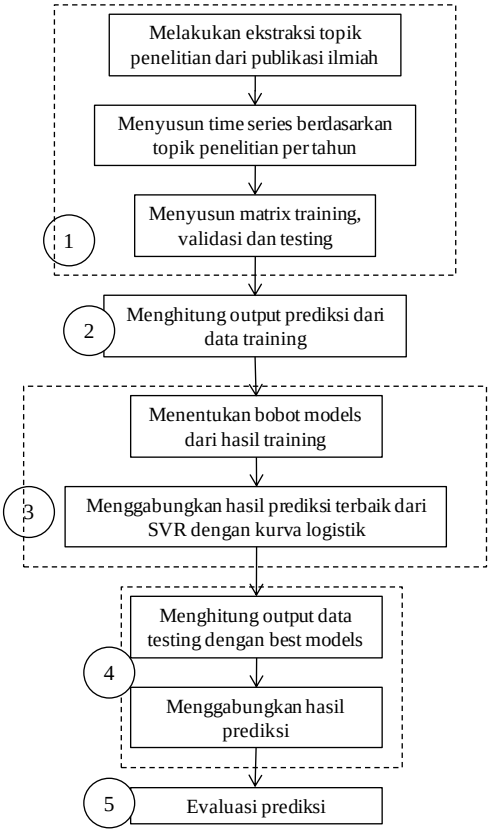
Sedangkan metode peringkat mirip dengan Borda count (Polikar, 2006), di mana setiap prediktor diberikan bobot berdasarkan posisinya dalam urutan. Misalnya, jika terdapat *N* buah posisi, kandidat pertama mendapatkan bobot *N-1*, kandidat kedua mendapatkan bobot *N-2*, dan yang ke-*i* mendapatkan bobot *N-i* suara. Kandidat terakhir menerima bobot 0.

4. METODOLOGI

4.1 Tahapan Penelitian

Langkah-langkah untuk melakukan percobaan, sebagaimana terlihat dalam Gambar 1, adalah: (1) menyusun data time series dari database hasil-hasil penelitian dan membangun matriks untuk pelatihan dan pengujian, (2) menjalankan algoritma prediksi, yang meliputi metode Support Vector Regression dengan beberpa parameter yang berbeda 3) memilih model terbaik dari data pelatihan, (4) menggabungkan hasil peramalan dengan dan kurva logistik, dan (5) mengevaluasi kinerja prediksi.

 ISSN: 1907-5022



Gambar 1. Tahapan Penelitian.

4.2 Data

Sumber data berasal dari data penelitian di Indonesia yang dikompilasi oleh situs Garuda (Garba Rujukan Digital) yang dikelola oleh Dirjen Pendidikan Tinggi, Kementerian Pendidikan Nasional. Topik yang diteliti sebanyak 14 buah, dengan pertimbangan ketersediaan data dengan rentang waktu minimal 16 tahun. Christodulous (2010) menyatakan bahwa minimum panjang training data adalah 16-20 titik.

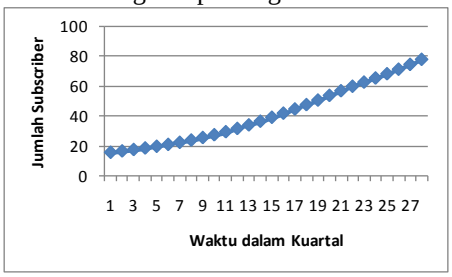
Tabel 1. Frekuensi topik per tahun dari Garuda.

Topik	**Tahun**				
	1995	**1996**	**1997**	**...**	**2010**
T_1	59	56	47		152
T_2	69	49	69		226
T_3	46	63	82		226
T_4	4	9	10		90
...					
T_{14}	2	7	6		22

Jumlah sample yang akan digunakan sebagai training dan testing dipengaruhi oleh panjang time series. Jika terdapat sebanyak nilai *k* untuk diprediksi, vektor y_{test} akan berisi nilai sebanyak *k*, dan matriks x_{test} terdiri dari *m*×*k*, dimana *m* adalah *sliding window*. Nilai *m* ditentukan ketika membentuk data untuk training, yakni x_{train} dan y_{train}, yang ukuran matriknya adalah *m*×*n* dan *n*. Semakin pendek ukuran *m*, semakin besar ukuran dataset, yakni *n*, yang bisa dibentuk, dan demikian pula sebaliknya.

Dengan data sebanyak 16 titik, dan titik yang akan diprediksi sebanyak 2 buah, maka terdapat 14 titik untuk training dan validasi. Dari 14 titik tersebut, disusun matrix 5×6 untuk x_{train} dan 1×6 y_{train}. Sedangkan untuk validasi masih disusun matrix 5×2 untuk x_{val} dan 1×2 y_{val}. Sedangkan dimensi matrix untuk testing serupa dengan dimensi matrix untuk validasi, tetapi dengan isi matrix yang berbeda.

Selain data dari Garuda, digunakan juga data dari ITU (*International Telecommunication Unit*) yang serupa dengan data yang digunakan oleh Christodulous (2010) sebagai acuan. Dalam tulisan tersebut, data yang digunakan frekuensi *World broadband penetration* dari tahun 1998 sampai 2008. Akan tetapi data yang paling serupa yang tersedia pada saat ini dari situs ITU adalah data *Mobile-cellular telephone subscriptions* setiap *quarter* dari tahun 2001 sampai 2010, dengan total sebanyak 28 titik. Dengan jumlah prediksi ke depan sebanyak 2 titik, penyusunan matrik training, validasi dan testing serupa dengan data sebelumnya.



Gambar 2. Data ITU.

4.3 Perangkat Penelitian

Ujicoba dalam penelitian ini menggunakan komputer dengan prosesor Pentium Core i3 dan memory sebesar 4GB. Software utama yang digunakan adalah Matlab versi 2008b. Data dinormalisasi dalam rentang -1 sampai 1, dengan perintah '*mapminmax*'. Toolbox tambahan untuk Support Vector Regression tersedia dari Gun (1998).

4.4 Evaluasi Kinerja

Untuk mengevaluasi kinerja, dihitung *Mean Squared Error* (MSE) terhadap data *out-of-sample*, yakni data yang tidak digunakan dalam training. MSE merupakan perbedaan antara nilai yang diestimasi dan nilai yang sesungguhnya. Jika $X=\{x_1, x_2,..x_T\}$ adalah suatu vektor input dalam domain *f*, dan nilai yang sebenarnya adalah $Y=\{y_1, y_2,..y_T\}$ maka untuk sample sebanyak *N*, tingkat kesalahan dihitung dengan

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 \qquad (6)$$

5. HASIL DAN PEMBAHASAN

5.1 Data Garuda

Kinerja SVR dengan dataset dari Garuda dapat dilihat pada Tabel 2. Dari tabel tersebut dapat disimpulkan bahwa fluktuasi dari time series cukup

besar, karena kernel RBF (*Radial Basis Function*) dengan *sigma* cukup besar, misal 5, yang memberikan kurva yang cenderung *smooth*, tidak sesuai dengan karakteristik kurva yang sebenarnya. Sebaliknya, dengan *sigma* yang lebih kecil memberikan MSE yang relatif lebih kecil pula. Kondisi kurva yang berfluktuasi ini juga dikuatkan oleh kernel Polynomial, dimana hasil terbaik diperoleh dengan *degree* yang besar.

Tabel 2. Kinerja SVR dari data Garuda.

Parameter SVR	**MSE**					
	ts1	**ts2**	**ts3**	**...**	**ts14**	**AVG**
RBF 0.01	0.85	0.01	0.14		0.28	0.39
RBF 0.05	0.85	0.01	0.14		0.28	0.39
RBF 0.1	0.85	0.01	0.14		0.28	0.39
RBF 0.5	0.73	0.01	0.14		0.32	**0.35**
RBF 1	2.44	0.01	0.14		1.61	0.60
RBF 2	5.38	0.06	0.08		371.81	27.94
RBF 5	7.15	1.90	0.52		1401.17	106.36
Poly 1	7.58	11.58	0.11		2774.16	385.49
Poly 2	6.20	0.09	0.17		295.22	22.28
Poly 3	5.48	0.01	0.12		37.95	3.91

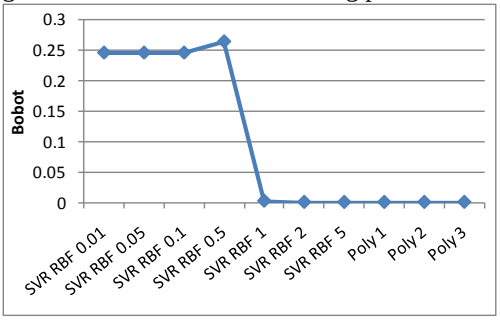
Catatan: ts adalah singkatan dari time series, AVG dari average.

Dari SVR dengan beberapa parameter yang berbeda tersebut, dilakukan penggabungan melalui rata-rata, pembobotan berdasarkan kinerja, dan pemeringkatan, yang hasilnya seperti terlihat pada Tabel 3. Dari tabel tersebut terlihat bahwa kombinasi dengan *inverse-MSE* memberikan hasil yang terbaik. Hasil kombinasi ini dapat memiliki tingkat kesalahan yang kecil, yakni 0.35, sama dengan error terkecil dari rata-rata MSE pada SVR terbaik pada Tabel 2.

Tabel 3. Kinerja Kombinasi SVR.

Metode Kombinasi	**MSE**					
	ts1	**ts2**	**ts3**	**...**	**ts14**	**AVG**
Avg	2.08	0.16	0.04		174.40	15.11
1/MSE	0.62	0.01	0.02		0.28	**0.35**
Rank	0.73	0.02	0.09		28.06	2.67

Sebagai ilustrasi, Gambar 3 memberikan besaran bobot dari tiap-tiap prediktor berdasarkan hasil MSE dari data validasi. Bobot yang tinggi pada gambar tersebut berkorelasi dengan tingkat MSE yang rendah berdasarkan data testing pada Tabel 2.



Gambar 3. Bobot *inverse-MSE* untuk kombinasi.

Langkah berikutnya adalah melakukan *curve fitting* dengan kurva logistik terhadap 14 titik dari data selain testing ditambah 2 data hasil prediksi terbaik. Hasil MSE setelah dilakukan kombinasi dengan kurva logistik ini adalah 0.32, sedikit lebih baik dari MSE sebelumnya.

Untuk mengetahui topik penelitian yang memiliki potensi cukup besar, atau biasa disebut dengan *emerging*, maka digunakan cara penghitungan *average year of publication* (Ziegler, 2009) seperti pada persamaan (4). Skor dari perhitungan tersebut, topik yang sedang berkembang adalah *Medicine*, *Biology* dan *Social pathology and public walfare*.

Tabel 4. Urutan topik yang *emerging* berdasarkan *average year of publications*.

Topik Penelitian	**Avg Year of Pub**
R Medicine (General)	15210.6
QH301 Biology	14900.3
HV Social pathology. Social and public welfare	14695.4
K Law (General)	13703.9
SF Animal culture	13490.0
QA Mathematics	12633.1
NA Architecture	12521.9
RA0421 Public health. Hygiene. Preventive Medicine	12156.6
SH Aquaculture. Fisheries. Angling	12072.0
TA Engineering (General). Civil engineering (General)	11926.2
QD Chemistry	11393.6
QC Physics	10195.8
P Philology. Linguistics	7267.3
S Agriculture (General); SF Animal culture	7159.5

5.2 Data ITU

Data ITU digunakan sebagai *benchmarking*, dimana peneliti terdahulu yang menggunakan kombinasi kurva logistik dan ARIMA mendapatkan MSE terbaik sebesar 0.025 (Chritodulous, 2010). Dengan data yang serupa, karena data yang sama tidak tersedia lagi pada situs ITU, penelitian ini menghasilkan MSE sebesar 0.0064.

Langkah yang dilakukan pada daset ini sama dengan langkah pada dataset sebelumnya, yakni menghitung MSE untuk tiap-tiap prediktor, mencari bobot dari tiap-tiap prediktor, menghitung prediksi dengan kombinasi, dan menggunakan hasil prediksi tersebut untuk dikalibrasi dengan kurva logistik.

Tabel 5. Kinerja SVR dari data ITU.

Parameter SVR	**MSE**
RBF 0.01	0.897
RBF 0.05	0.897
RBF 0.1	0.897
RBF 0.5	0.791
RBF 1	0.419
RBF 2	0.121
RBF 5	**0.008**
Poly 1	**0.002**
Poly 2	0.144
Poly 3	0.195

Hasil MSE tiap-tiap prediktor dapat dilihat pada Tabel 5, yang mengindikasikan bahwa kurva

relatif *flat* karena MSE yang baik (atau rendah) dihasilkan oleh SVR dengan kernel Polynomial degree rendah (linear), yakni 1, atau SVR dengan kernel RBF dengan sigma cukup besar, yakni 5.

Kombinasi SVR dengan parameter yang berbeda dengan inverse-MSE menghasilkan MSE sebesar 0.007, yang kemudian diperbaiki lagi dengan kurva logistik menjadi 0.0064. Sebagai catatan, meskipun secara individu, kenel polynomial dapat menghasilkan MSE 0.002, akan tetapi untuk data yang berbeda, kernel ini tidak akan selalu lebih baik dari kernel yang lain, sebagaimana diperlihatkan oleh Tabel 2. Oleh karena itu, tetap diperlukan mekanisme kombinasi prediksi.

6. KESIMPULAN

Dari hasil uji coba dapat disimpulkan bahwa pendekatan kurva logistik dapat digabungkan dengan metode *machine learning* untuk data yang memiliki sifat pertumbuhan seperti kurva-S. Akan tetapi, dataset yang lebih beragam dan lebih panjang rentang waktunya diperlukan untuk memberikan penegasan akan kebenaran hipotesa ini.

Meskipun dataset dari Garuda yang digunakan sudah memenuhi syarat minimal panjang time series, tetapi jumlah record dalam matrix training dan validasi yang dihasilkan masih relatif kecil untuk melakukan generalisasi. Disamping itu, dataset ini memiliki fluktuasi yang cukup besar sehingga relatif lebih sulit untuk dihasilkan pola dari jumlah training yang terbatas.

Untuk penelitian lebih lanjut, selain penambahan dataset, perlu juga dicoba untuk menggunakan mekanisme lain dalam kombinasi prediksi, misalnya menggunakan kernel learning. Jenis topik penelitian juga dapat di-*generate* dari kelompok frase, misalnya dari judul, abastrak ataupun deskripsi lainnya.

PUSTAKA

Bengisu M, Nekhili R.. (2006). Forecasting emerging technologies with the aid of science and technology databases. *Technological Forecasting & Social Change* 73 (2006) 835–844.

Christodoulos C., Michalakelis C., Varoutas D. (2010). Forecasting with limited data: Combining ARIMA and diffusion models. *Technological Forecasting & Social Change* 77 (2010) 558–565.

Cristianini, N.. (2001). Support Vector and Kernel Machines. *ICML (International Conference on Machine Learning)*.

Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8):981–1012, October 2006.

Gunn S. R. (1998). Support Vector Machines for Classification and Regression. *Technical Report*, University Of Southampton, 10 May 1998.

Jun, S., Uhm, D. (2010). Technology Forecasting Using Frequency Time Series Model: Bio-Technology Patent Analysis. *Journal of Modern Mathematics & Statistics*, 4(3):101-104.

Kobayashi, S., Shirai Y., Hiyane K., Kumeno F., Inujima H. dan N. (2005). Technology Trends Analysis from the Internet Resources. *PAKDD 2005*, LNAI 3518, pp. 820–825, 2005, Springer-Verlag Berlin Heidelberg.

Mathews, J. H. (1992). Bounded Population Growth: A Curve Fitting Lesson. *Mathematics and Computer Education*, California State Univ.

Muller, K. R., Smola, A. J., Ratsch, G., Scholkopf, B., Kohlmorgen, J., and Vapnik, V. (1997). Predicting Time Series with Support Vector Machines. *Proceedings of the 7th International Conference on Artificial Neural Networks*, Springer-Verlag London, UK.

Phienthrakul T., Kijsirikul B., (2006). Evolutionary Support Vector Regression based on Multi-Scale Radial Basis Function Kernel, *NN3 Forecasting*.

Polikar, R. (2006). Ensemble Based System in Decision Making. *IEEE Circuits And Systems Magazine*.

Rahayu, E. R, and Hasibuan, Z. A. (2006). Identification of technology trend on Indonesian patent documents and research reports on chemistry and metallurgy fields. *Proceeding Asia Pacific Conf.*, Singapore.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, 68(3):595–610.

Vidican, G., Woon, W. L., Madnick, S. (2009). Measuring Innovation Using Bibliometric Techniques: The Case of Solar Photovoltaic Industry, *Working Paper CISL# 2009-05*.

Widodo, A., Fanany, M. I., Budi I. (2011). Technology Forecasting in the Field of Apnea from Online Publications: Time Series Analysis on Latent Semantic. *International Conference on Digital Information Management*, 26-28 September 2011, Melbourne, Australia.

Woon, W.L., Hensche, A., and Madnick, S. (2009). A Framework For Technology Forecasting And Visualization. *Working Paper Series*, ESD-WP-2009-16, October 2009.

Zhu, D., Porter A., Cunningham, S., Carlisie, J., Nayak A. (1998). A process for mining science and technology document databases, illutrated for the case of 'knowledge discovery and data mining'. *Technology Policy and Assessment Center*, Georgia Institute of Technology, Atlanta, GA.

Ziegler, B. E. (2009). Methods for Bibliometric Analysis of Research: Renewable Energy Case Study. *Working Paper CISL#2009-10*, September 2009.