非定形文書中の日程情報を自動配信するスケジュールリマインダ

2H－1

池田 崇博　佐藤 研治　奥村 明俊
NEC C&C メディア研究所

1　はじめに

イントラネット・インターネットの普及により、電子メイルや Web ページ等の非定形文書が広く流通するようになった。こうした中、非定形文書から意味のある情報を抽出し、文書を有効に活用していくことが重要になってきている。筆者らは、これまでに、文書中から 5W1H (いつ、どこで、だれが、なにを、どうした) 情報を抽出し、文書検索に活用する手法を提案してきた [1]。

イントラネットでにおける業務に関する電子メイルでは、会議開催通知や納期の指定にように、5W1H の中でも、日付・時間等の When を表す情報が重要な意味を持っていることが多い。しかしながら、そのような電子メイルを保存しておいたとしても、雑多なメイルの中に埋もれてしまい、後日その日付・日時を探し出すのに苦労することが多い。結局スケジュールを忘れないようにするためには、スケジュールを管理するためのソフトウェア (スケジューラ) を使い、メイルで届いた通知や指示からスケジュールに関する情報を手作業で拾い出し、改めてスケジューラに登録するという手間をかけているのが現状である。

同様のことはインターネットの Web 上の情報についても当てはまる。例えば、イベントの開催日程の情報などが掲載されている Web ページをブラウザにブックマークとして登録しておくことができるが、ブックマークだけでは、そのイベントを思い出すきっかけにはなりにくい。よいタイミングでブックマークを参照しないと、そのイベントについては忘れがちであるため、結局手間をかけてスケジューラに登録することになる。また、Web の場合、新しい情報はこちらから取りにいかない限り得られないため、頻繁にそのページをチェックしていなければ新しい情報を得られないという問題もある。

このような問題点を解決するシステムとして、本稿では、電子メイルや Web 上の情報から期日や期限などのスケジュールに関する情報を自動的に抽出し、その日時の前にユーザに通知するスケジュールリマインダシステムを提案する。

スケジュール情報は、電子メイル等の非定形文書からでも、パターンマッチングにより高い精度での情報の抽出が可能であると報告されている [2]。しかしながら、スケジュールの抽出漏れや抽出誤りが少しでもあると、重要なスケジュールを逃す可能性があり、実際にスケジュールリマインダとして用いることは難しい。そこで、本稿では、すべてのスケジュール情報に含まれ、頑健な抽出が可能な日時表現に着目してスケジュール情報を抽出する。

Schedule Reminder for Schedules in Free Form Documents
Takahiro Ikeda, kenji Satoh, and Akitoshi Okumura
NEC C&C Media Research Labolatories

2　スケジュール情報の抽出

ユーザにとって必要なスケジュール情報としては、そのスケジュールの日時についての情報のほかに、各スケジュールの内容ごとに固有な詳細情報とがある。例えば、会議開催のスケジュールについての情報であれば、会議の開催場所・会議の議題・会議の参加者といったものが詳細情報に当たる。提出物の〆切についての情報であれば、提出物の内容・提出物の提出先といったものが詳細情報である。

これらのうち、日時はすべてのスケジュール情報に必ず存在する情報であり、パターンとしての記述が可能であるため、パターンマッチングにより頑健な抽出が可能である。本システムでは、電子メイルや Web 上の文書で通常用いられている記述フォーマットを分析し、日時表現を日本語・英語合わせて 20 種類のパターンで表現した。

こうして、すべての日時表現から日時を抽出するが、実際にはそのすべてがスケジュールを表しているとは限らない。また、Web 上で公開されている情報にはさまざまなものがあるが、ユーザがあらゆる種類のスケジュールについていつも通知して欲しいとは限らない。そこで、本システムでは、日時表現の近傍に出現する文字列によりスケジュール情報の種類を定義し、この種類ごとにユーザがそのスケジュール情報を通知してもらうかどうかを指定できるようにする。

例えば、日時表現と同一の行に「締切」・「締め切り」・「〆切」のいずれかが出現するとき、その日時表現が表すスケジュール情報の種類を締切と定義する。その上で、種類が締切のものだけを通知するように指定しておくことで、ユーザが締切の情報だけを通知してもらうことができるようにする。

各スケジュールの詳細情報についても、頑健な抽出を実現するために、日時表現の前後の数行すべてを、その日時に関するスケジュールの詳細情報として抽出するようにする。これは、実際の電子メイル等において、1つの事柄に関する記述は、狭い範囲に集中して現れるという観察に基づいている。

3　スケジュールリマインダシステムの構成と動作

本システムを、図 1 に示すような構成で Windows95/NT 上に実装した。

ユーザは、予めパラメータ設定モジュールを通して、スケジュールを抽出するメイルのあるフォルダ、スケジュール情報が掲載されている WWW サイトの URL、抽出するスケジュール情報の種類、各スケジュールを何日前から通知するか、日時表現の前後何行をスケジュールとして抽出するか等のパラメータを設定する。

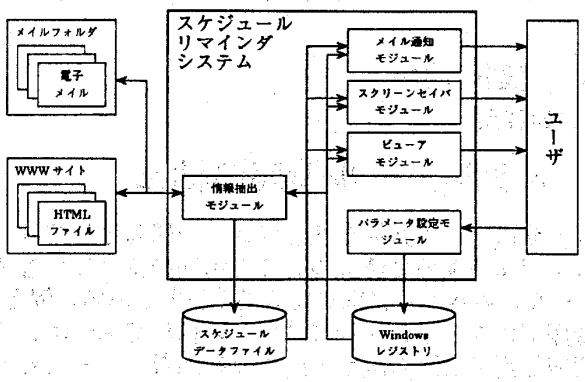

図 1: スケジュールリマインダシステムの構成

情報抽出モジュールは、定期的に起動され、指定された電子メイルのフォルダや指定された URL から文書を収集し、前節の方式に従ってスケジュールを抽出する。

メイル通知モジュールは、抽出されたスケジュール情報のうち、通知すべき期日になったものを選んで電子メイルによりに通知する。ユーザは、スケジュールビューアを立ち上げることにより、能動的にスケジュールを確認することもできる。また、スクリーンセイバとして、ユーザが作業を中断したときにスケジュールを画面に流すことができるようになっている。

図 2 にスケジュールビューアの画面を示す。メイルで届いた論文提出締切の情報や WWW 上で公開されている研究会のスケジュール等が抽出され、7 月 31 日のスケジュールとして表示されている。ユーザにメイルで通知されるスケジュールも、ほぼこれと同じ情報である。

4 考察

実際に使用しているメイルフォルダ中の約 2,200 件のメイル、および、情報処理学会のカレンダーページやコンピュータ関連のイベント情報等の 12 件の URL を対象として、本システムを実際に試用した結果、日時表現に関しては、ほぼ漏れなく抽出できており、メイルや Web ページなどの非定形文書からスケジュールを抜き出して通知するツールとして、十分に機能していることが分かった。

現在のスケジュール抽出の方法では、「明日」「来週」などの相対表現に対応していないため、若干の抽出漏れが存在する。実際に相対表現が使われるのは、スケジュールとして管理するまでもない、ごく近い期日である場合が多いため、これはあまり問題とはならない。しかし、電子メイルには、それが発信された日時が記録されているため、その情報を利用して相対指定された日時を特定することも検討している。

逆に、分数の 1/2 を 1 月 2 日と認識する、あるいは、相手の都合を尋ねている「4/20 か 4/21 はいかがでしょうか?」というような場合でも 4 月 20、21 日のスケジュールとして認識するといったように、スケジュールとして余分なものを抽出してしまう場合もある。本システムでは、日時表現に着目してスケジュールの可能性があるものは漏れなく抽出し、抽出されたもののうち、近傍に適切なキーワードがあるものだけを残すことで抽出精度を高めている。ユーザが必要なスケジュール情報に合わせてキーワードを設定することで、各ユーザに合わせてより精度の高い抽出を実現する仕組みである。

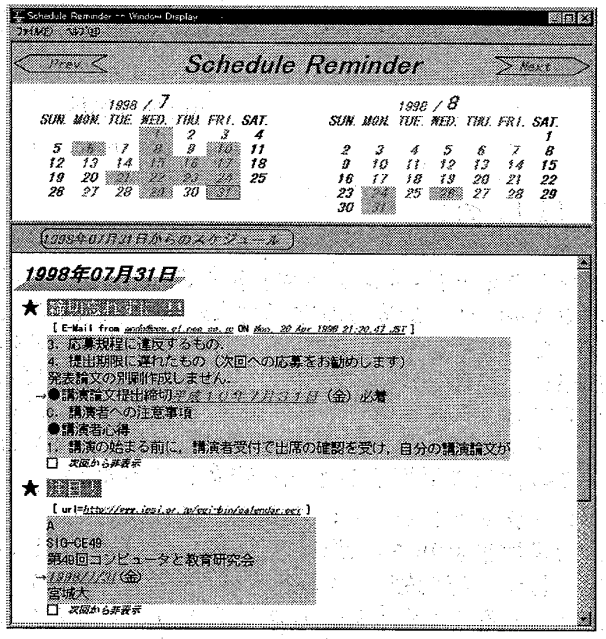

図 2: スケジュールビューアの画面

5 まとめ

本稿で提案するシステムにより、電子メイルや Web ページ等の非定形な文書から、期日や期限などのスケジュールに関する情報を自動的に抽出し、その日時の前にユーザに通知することが可能になった。

今後は、学会の締切の情報に対して学会発表の申し込み手順を提示するなどのように、スケジュール情報の種類ごとにそのスケジュールの処理に必要な情報をセットで提供する機能や、スケジュールの時間についても抽出し、スケジュールをワンタッチで既存のスケジュール管理ソフトに登録できるようにする機能等を実現するなどの改良を行っていく。

参考文献

[1] 池田崇博, 奥村明俊, 村木一至, "MIIDAS: 情報の選別と Easy Reading のためのエピソード," 情報処理学会第 55 回全国大会, 5Q-10 (1997).

[2] 長谷川隆明, 高木伸一郎, "電子メールコミュニケーションにおけるスケジュール情報の抽出," 情報処理学会第 123 回自然言語処理研究会 (1998).