Journal Of Electronical & Cyber Defence
Vol. 2, No. 4, 2015, Serial No. 8

Online Collaborative Planning in Complex Environment

S. Sadati[1*], M. NaghianFesharaki[2], A. H. MomeniAzandariani[3]

1-Master Student, Malek Ashtar University of Technology

2- Associated professor, Malek Ashtar University of Technology

3- PhD Student, Malek Ashtar University of Technology
(Received: 29/09/2013 , Accepted: 11/05/2015)

ABSTRACT

Although existing Planning methods can plan under uncertainty and decentralize situation, most of them malfunction in some complicated conditions of command and control scenarios such as real time decision making, need accurate planning, bounded communication between agents, dynamic worlds and partially observable environments. Among suitable models for these situations, we can consider extended models of DEC-POMDPs such as MAOP-COMM that can handle these conditions. It is possible to improve MAOP-COMM model to do planning for agents with double precision. In this paper we have improved the algorithm of MAOP-COMM model by upgrading value function heuristic and using "two steps lookahead" in the strategy of finding best policy and making correct decision. Improved algorithm performs online planning for agents in a multi agent system in uncertain condition with better performance and high percent of correct decision making. We experiment resulted algorithm on Grid Soccer benchmark. The results obtained prove efficiently of proposed improvements.

Keywords: Online Collaborative Planning, MultiAgent Systems, Decentralize POMDPs, Policy Tree, Decision Theory.

* Corresponding Author Email: sadati_saeedeh@yahoo.com

مجله علمی- پژوهشی "پدافند الکترونیکی و سایبری"
سال دوم، شماره ۴، زمستان ۱۳۹۳، ص ۲۳-۱۵

طرح ریزی مشارکتی بر خط در محیط‌های پیچیده

سعیده ساداتی[1*]، مهدی نقیان فشارکی[2]، امیرحسین مومنی ازندریانی[3]
۱- دانشجوی کارشناسی ارشد هوش مصنوعی، دانشگاه مالک اشتر
۲- دانشیار،دانشگاه مالک اشتر ۳- مربی، دانشگاه مالک اشتر
(دریافت: ۹۲/۷/۷؛ پذیرش: ۹٤/۲/۲۱)

چکیده

رویکردهای طرح‌ریزی موجود اگرچه می‌توانند در شرایط عدم قطعیت و به‌صورت غیرمتمرکز طرح‌ریزی نمایند، اما اکثر آن‌ها در مواقعی که شرایط پیچیده سناریوهای فرماندهی و کنترل همچون نیاز به طرح‌ریزی دقیق، تصمیم‌گیری بلادرنگ، ارتباطات محدود بین عامل‌ها و پویایی محیط حاکم است ، بازده خوبی نداشته و گاهی با شکست مواجه می‌شوند. از بین رویکردهای موجود، مدل‌های توسعه‌یافته DEC-POMD مانند MAOP-COMM، برای این شرایط مناسب هستند. می‌توان با بهبود مدل MAOP-COMM، طرح‌ریزی دقیق‌تری برای عامل‌ها انجام داد. ما در این مقاله با ارتقاء الگوریتم اکتشافی تابع ارزش و به‌کارگیری پیش‌بینی دو مرحله‌ای در راهبرد یافتن سیاست بهینه و اخذ تصمیم صحیح، الگوریتم اخیر را بهبود داده‌ایم. الگوریتم بهبودیافته پیشنهادی در شرایط عدم قطعیت به صورت غیر متمرکز و برخط با کارایی مضاعف و درصد بالایی از صحت تصمیم‌گیری برای عامل‌ها طرح‌ریزی می‌کند. الگوریتم حاصل روی بنچ مارک Grid Soccer آزمایش شده است. نتایج به‌دست آمده، برتری الگوریتم ارائه‌شده با بهبودهای پیشنهادی را نشان می‌دهد.

واژه‌های کلیدی: طرح‌ریزی مشارکتی بر خط، سیستم‌های چند عاملی،POMDP غیرمتمرکز، درخت سیاست، تئوری تصمیم‌گیری.

۱- مقدمه

طرح‌ریزی عبارت است از رشته اقداماتی که به منظور رسیدن به یک هدف انجام می‌شوند. در علم کامپیوتر، طرح‌ریزی چندعاملی عبارت است از همگرا کردن منابع و فعالیت‌های چندین عامل، یا به عبارت دیگر، طرح‌ریزی همراه با مشارکت [۱].

یک سیستم چندعاملی (MAS)[1] متشکل از چندین عامل مستقل از هم است که در یک فضا برهم اثر متقابل دارند. هر عامل یک تصمیم‌گیرنده است که در محیط واقع شده است و به طور مستقل بر اساس مشاهدات و دانش حوزه خودش اقدام می‌کند تا به هدف معینی دست یابد. در یک سیستم چندعاملی، عامل‌ها می‌توانند اهداف متفاوت و حتی متضاد داشته باشند. طراحی سیستم چندعاملی برای بسیاری از حوزه‌های هوش مصنوعی مفید است؛ به ویژه وقتی که یک سیستم از چندین موجودیت تشکیل شده باشد که به‌صورت خاص یا بر اساس وظیفه توزیع شده‌اند.

در محیط‌های چندعاملی هر عامل باید محدودیت ناشی از اقدامات عامل‌های دیگر را نیز در طرح ریزی در نظر بگیرد. از این رو فرایند طرح‌ریزی چندعاملی پیچیده‌تر از طرح‌ریزی تک‌عاملی است. طرح‌ریزی چند عاملی می‌تواند متمرکز یا غیر متمرکز[2] باشد. منظور ما از طرح‌ریزی مشارکتی، طرح‌ریزی چند عاملی غیرمتمرکز است. در طرح‌ریزی غیرمتمرکز هر عامل به‌طور مستقل برای خودش تصمیم می‌گیرد که بر اساس اطلاعاتی که دریافت کرده یا مربوط به او است، چه اقدامی انجام دهد یا چه طرحی را پیاده‌سازی کند [۲].

محیط‌های پیچیده، ویژگی‌ها و پارامترهای خاص خود را دارند؛ از جمله مهم‌ترین این پارامترها می‌توان به عدم قطعیت، پویایی، نیمه‌رویت‌پذیری و نقص اطلاعات، تنوع در وظایف عامل‌ها[3]، بلادرنگ بودن و نیاز به حذف مدیریت مرکزی اشاره کرد [۷]. از این جهت

* رایانامه نویسنده مسئول: sadati_saeedeh@yahoo.com
1- Multi Agent System
2- Decentralize
3- Diversity

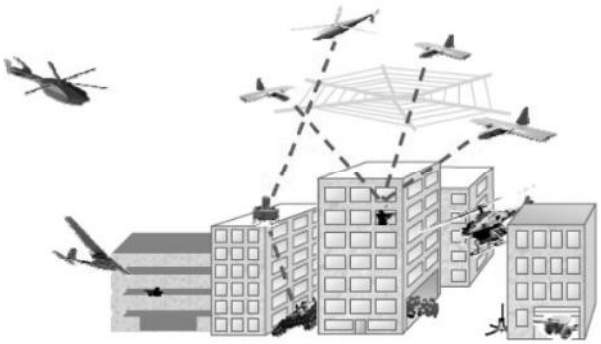
طرح‌ریزی در این محیط‌ها با دشواری‌هایی همراه است و برای انجام یک طرح‌ریزی دقیق در سناریوهای سیستم‌های چندعاملی محیط‌های پیچیده، نیازمند در نظر گرفتن پارامترهای مختص این محیط‌ها هستیم [۱۴].

فرایند طرح‌ریزی مشارکتی در حوزه‌های دارای اهداف رقابتی، تصمیم‌گیری غیرمتمرکز یا کنترل توزیع‌شده که دارای محدودیت‌های محاسباتی و یا ارتباطی هستند، به کار می‌رود. این حوزه‌ها بازه وسیعی از کاربردهای صنعتی، شبکه، لجستیک، نظامی، بازی‌ها و سایر زمینه‌ها را دربر می‌گیرند. نمونه‌هایی از هریک از این کاربردها عبارت اند :

- حوزه نظامی: همکاری و هماهنگی زیرسامانه‌های دفاعی نظیر واحدهای توپخانه، شبیه‌سازهای صحنه نبرد، حسگر های توزیع‌شده، تجهیزات بدون سرنشین[1]، سامان‌دهی دارایی ها و نیروها، طرح‌ریزی خودکار حمله و مانور .
- حوزه لجستیک: مدیریت زنجیره‌های توزیع[2]، مدیریت خرید و فروش و هماهنگی با خریداران و فروشندگان، برنامه‌ریزی GIS[3]
- حوزه صنعت: مونتاژ ماشین، مدیریت کارخانه، مدیریت نیروی کار [۱۲].
- حوزه شبکه: کنترل غیرمتمرکز شبکه‌های حسگر بی سیم، طرح‌ریزی جهت هماهنگ نگه‌داشتن عامل‌های نرم‌افزاری در فضای سایبر .
- حوزه تحقیقات فضایی: سنجش از راه دور، کنترل پهپادها، ماهواره‌ها و فضاپیماها.
- حوزه بازی‌های رایانه‌ای: طرح‌ریزی مشارکتی برای روبات‌های بازیگر تیمی مانند روبات‌های فوتبالیست.

شکل (۱)، نمونه‌ای از یک سناریوی نظامی که با استفاده از طرح‌ریزی مشارکتی مدیریت و کنترل می‌شود را نشان می‌دهد.

طرح ریزی در دنیای چندعاملی را می‌توان به روش‌های مختلفی مدل کرد. از جمله این روش‌ها می‌توان عامل‌های شناختی BDI، تئوری بازی‌ها و فرآیندهای تصمیم مارکوف (MDP)[4] و مدل‌های توسعه‌یافته آن از جمله POMDP، DEC-POMDP[5] را نام برد. از بین مدل‌های طرح‌ریزی موجود، رویکردهای توسعه‌یافته مدل DEC-POMDP مانند MAOP-COMM با پارامترهای یادشده انطباق نسبی دارند و می‌توانند تیمی از عامل‌های همکار را که با یک محیط تصادفی نیمه‌رویت‌پذیر در ارتباط هستند مدل کنند [۱۳]. در سال‌های اخیر چندین توسعه بر این رویکردها انجام شده است که بیشتر آن‌ها غیر برخط هستند؛ یعنی قبل از اجرا بهترین اقدام را برای اجرا در تمام وضعیت‌ها محاسبه می‌کنند. از جمله جدیدترین توسعه برخط روش فرآیندهای تصمیم‌گیری مارکوف نیمه‌رویت پذیر غیرمتمرکز می توان به مدل MAOP-COMM اشاره کرد. الگوریتم‌های برخط، با دریافت تمام اطلاعات موجود در هر لحظه فقط یک گام طی می‌کنند. MAOP-COMM الگوریتمی است که با محدود کردن ارتباطات عامل‌ها و استفاده از مدل ارتباطی sync هماهنگی بین عامل‌ها را در طرح‌ریزی برخط تضمین می‌کند. اما این الگوریتم به دلیل ضعف تابع اکتشافی[6] و استفاده از پیش‌بینی یک مرحله‌ای بازده خوبی در اتخاذ سیاست بهینه و انتخاب اقدامات برگزیده ندارد. در این مقاله سعی شده است تا با ارتقاء تابع اکتشاف و افزایش گام‌های پیش‌بینی، بهره بیشتری در اتخاذ سیاست بهینه و انتخاب اقدامات برگزیده حاصل شود.



شکل (۱). نمونه‌ای از طرح‌ریزی چندعاملی برای حمله و مانور نظامی [۱۲].

۲- مدل رسمی POMDP غیرمتمرکز

مدل POMDP غیرمتمرکز یک توسعه طبیعی از چارچوب زنجیره تصمیم مارکوف را برای شرایط چندعاملی و مشارکتی پیشنهاد می‌دهد که برای برنامه‌ریزی و یادگیری در شرایط عدم قطعیت بسیار مفید است. یک فرآیند تصمیم‌گیری مارکوف نیمه‌رؤیت پذیر غیرمتمرکز یک چندتایی به‌صورت زیر است:

$<I, S,\{Ai\}, \{\Omega i\},P, O, R, b0>$

1- Unmanned
2- Supply Chain Management
3- Geographic Information System
4- Markov Decision processes

5- Decentralize Partially Observable Markov Decision processes
6- heuristic

- I مجموعه متناهی از عامل‌ها با اندیس‌های ۱،... ،n است. وقتی n=1 یک DEC-POMDP با یک POMDP تک‌عاملی برابر است.

- S یک مجموعه متناهی از حالت‌های سیستم است. یک حالت تمام ویژگی‌های مربوط به سیستم پویا و ویژگی‌های مارکوف را دارا است. به عبارت دیگر، احتمال حالت بعد فقط به حالت جاری و اقدام مشترک وابسته است نه به حالت‌های قبلی و اقدامات مشترک:

$$P(st+1|s^0,a^0, \dots , s^{t-1}, a^{t-1}, s^t, a^t) = P(s^{t+1}|s^t,a^t).$$

- A_i مجموعه محدودی از اقدامات در دسترس برای عامل i است و $A=X_i \times A_i$ مجموعه اقدامات مشترک است که $a=< a_1,..., a_n >$ یک اقدام مشترک را نشان می‌دهد. ما فرض می‌کنیم عامل‌ها نمی‌بینند که چه اقداماتی به وسیله سایر عامل‌ها در هر گام زمانی اتخاذ شده است.

- Ω_i مجموعه متناهی از مشاهدات در دسترس برای عامل i است و $\Omega =X_i \times \Omega_i$ مجموعه مشاهدات است که $o=<o_1, \dots , o_n>$ یک مشاهده اشتراکی در لحظه است. در هر گام زمانی، محیط فقط یک مشاهده اشتراکی از خود نشان می‌دهد، اما هر عامل فقط قسمت خود را مشاهده می‌کند.

- P یک جدول احتمال انتقال حالت است. $P(s'|s, a)$ نشان‌دهنده احتمال اتخاذ اقدام a در جابه‌جایی از حالت s به حالت ′s است. P تأثیر ناگهانی اقدامات را بر روی محیط توصیف می‌کند. فرض برآن است که احتمال جابه‌جایی بدون تغییر است، بدان معنا که این احتمالات مستقل از گام‌های زمانی هستند.

- O جدول احتمال مشاهدات است. $O(o|s', a)$ نشان‌دهنده احتمال دیدن مشاهده اشتراکی o بعد از اقدام مشترک a و رسیدن به حالت ′s است. O توصیف‌گر آن است که عامل‌ها چگونه حالت محیط را دریافت می‌کنند. احتمال مشاهده نیز ایستا در نظر گرفته می‌شود.

- $R: S*A \rightarrow \mathbb{R}$ تابع بهره است. $R(s, a)$ نشان‌دهنده بهره به‌دست آمده از اتخاذ اقدام مشترک a در حالت s است.

- $b^0 \in \Delta(S)$ توزیع حالت گمان اولیه است. b^0 برداری است که یک توزیع احتمال گسسته را روی S (که دانش کلی عامل را درباره حالت اولیه دربردارد)، معین می‌کند.

- h_i یعنی پیشینه عامل i به عنوان دنباله‌ای از اقدامات اتخاذشده و مشاهدات دریافت‌شده توسط عامل i تعریف می‌شود [۵]. در هر گام زمانی t، $h_i^t = (a^0_i,o^1_i, \dots , o^{t-1}_i, a^{t-1}_i, o^t_i)$ پیشینه عامل i است و $h^t = <h_1^t, ..., h_n^t>$ یک پیشینه مشترک است. اصطلاح باور مشترک $b(.|h) \in \Delta(S)$ توزیع احتمال روی حالاتی است که در پیشینه h واقع شده‌اند. اگر مجموعه‌ای از پیشینه‌های مشترک از گام پیش داشته باشیم، با استفاده از قاعده بیز می‌توان مجموعه‌ای از حالت‌های باور مشترک گام جاری را محاسبه کرد:

$$\forall s' \in S , b^t (s'|h^t) = \quad (1)$$

$$\frac{O(o^t \mid s', a^{t-1}) \sum_{s \in S} P\left(s' \mid s, a^{t-1}\right) b^{t-1}(s \mid h^{t-1})}{\sum_{s'' \in S} O(o^t \mid s'', a^{t-1}) \sum_{s \in S} P\left(s'' \mid s, a^{t-1}\right) b^{t-1}(s \mid h^{t-1})}$$

۲-۱- مقایسه الگوریتم‌های برخط و غیر برخط

در سال‌های اخیر، چندین توسعه بر مسائل DEC-POMDP انجام شده است که بیشتر این الگوریتم‌ها غیر برخط هستند؛ یعنی قبل از اجرا بهترین اقدامات را برای اجرا در تمام وضعیت‌ها محاسبه می‌کنند. اگرچه این الگوریتم‌های غیر برخط می‌توانند به بازده خیلی خوبی دست یابند، اما اغلب، فضای روشی که می‌پیمایند به صورت نمایی مضاعف می‌شود؛ از این رو زمان زیادی را صرف می‌کنند. مثلاً BPIP-IPG که جدیدترین تکنولوژی مبتنی بر الگوریتم غیربرخط MBDP است، ۸۵ ساعت به طول می‌انجامد تا مسئله کوچکی مانند ملاقات در یک شبکه[1] ۳×۳ را که شامل ۸۱ حالت، ۵ اقدام و ۹ مشاهده است حل کند[۳]. در طرح‌ریزی غیربرخط، طرح‌های محاسبه‌شده بین عامل‌ها توزیع و به‌وسیله هر عامل بر اساس اطلاعات محلی‌اش اجرا می‌شود. بنابراین با اینکه اجرا غیرمتمرکز است، فاز طرح‌ریزی متمرکز است [۲].

الگوریتم‌های برخط، علی‌رغم الگوریتم‌های غیر برخط، وضعیت‌های غیر منتظره و پیش‌بینی نشده را بهتر اداره می‌کنند. الگوریتم‌های طرح‌ریزی برخط، اغلب طرح‌ریزی را با اجرا درمی‌آمیزند؛ آن‌ها به‌جای این که مانند الگوریتم‌های غیربرخط همه طرح را ایجاد کنند، فقط نیاز به یافتن اقدامات گام جاری دارند و با دریافت تمام اطلاعات موجود جاری، در هر لحظه فقط یک گام طی می‌کنند. توسعه‌های اخیر در الگوریتم‌های برخط اظهار می‌کنند که

1- Meeting in a Grid

ترکیب تکنیک‌های برخط با ارتباطات منتخب- هنگامی که ارتباطات ممکن باشد- می‌توانند مؤثرترین راه غلبه بر مسائل DEC-POMDP بزرگ باشند. از جمله جدیدترین توسعه برخط روش فرآیندهای تصمیم‌گیری مارکوف نیمه‌رویت‌پذیر غیرمتمرکز می‌توان به مدل MAOP-COMM اشاره کرد.

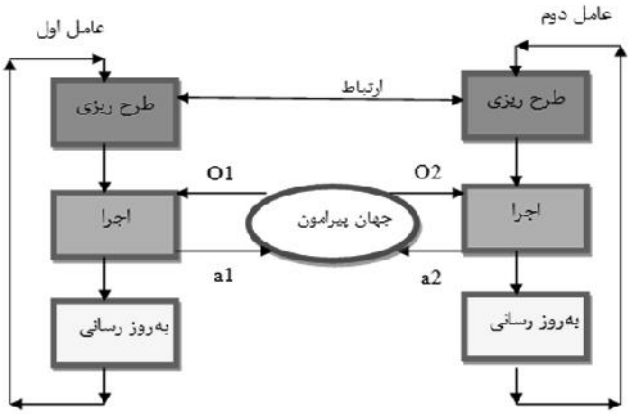
۲- ۲- مدل MAOP-COMM

این الگوریتم یکی از جدیدترین توسعه‌های مدل DEC-POMDP به‌شمار می‌رود و همچون مدل والد خود توانایی مدل‌سازی طرح‌ریزی در محیط‌های پیچیده را دارد چرا که از پارامترهایی چون عدم قطعیت، کانال ارتباطی محدود (نیاز به ارتباطات محدود)، و میدان دید محدود (نیمه‌رویت‌پذیر) پشتیبانی می‌کند. طرح‌ریزی طولانی‌مدت به صورت برخط دشوار است. این مدل برای غلبه بر این چالش، ارتباطات بین عامل‌ها را محدود کرده و برای ارتباط از مدل ارتباطی sync بهره برده است [۲]. در این روش هر عامل به‌صورت مستقل طرح‌ریزی نموده و بر اساس مشاهدات محلی خود انبار باوری می‌سازد و هرگاه مشاهدات جدید وی با انبار باور در تناقض باشد، با عامل همکار خود ارتباط برقرار می‌کند. در پیاده‌سازی انبار باور، هر باور برای یک عامل با استفاده از فرمول شماره(۱)، برای هر پیشینه محاسبه می‌شود.

الگوریتم MAOP به‌صورت موازی به‌وسیله تمام عامل‌های تیم اجرا می‌شود که در آن، طرح‌ریزی و اجرا گنجانده شده است. به‌طور دقیق‌تر، این الگوریتم برخط به فاز طرح‌ریزی، فاز اجرا و فاز به‌روزرسانی تقسیم می‌شود که به‌صورت متوالی در هر گام زمانی اجرا می‌شوند. در شکل (۲) ساختار این الگوریتم نشان داده شده است. در الگوریتم MAOP هر عامل به صورت مستقل طرح‌ریزی می‌کند تا اقدام مناسب برای لحظه کنونی را بیابد. عامل اقدام را در فاز طرح‌ریزی انجام می‌دهد و در نتیجه به حالت جدیدی از سناریو می‌رود. در حالت جدید مشاهده دریافتی از محیط را ذخیره می‌کند تا در طرح‌ریزی گام بعد آن را به‌کار برد. در مرحله بعد باورها به‌روز رسانی می‌شوند و مجدداً عامل وارد فاز طرح‌ریزی می‌شود. چنانچه مشاهده دریافتی عامل با باورهایش ناسازگار باشد، عامل اقدام به ارتباط با عامل همکار خود می‌کند.

مشاهده و اقدام در ساختار این مدل را می‌توان به مدل حلقه کنترلی بوید (OODA) شبیه کرد. این مدل در سال ۱۹۸۷به‌منظور بازنمایی سازوکار تصمیم‌یار جهت سیستم‌های نظامی ارائه شد. این مدل در داده‌آمیزی به نحو گسترده‌ای مورد استفاده قرار می‌گرفت [۸].

تفاوت اصلی مدل OODA[1] با مدل مذکور در این است که OODA برای سیستم‌های تصمیم‌یار که تعامل انسان و ماشین مطرح است به‌کار می‌رود اما MAOP، مدل تصمیم‌گیری عامل‌ها مستقل از دخالت انسان است.



شکل (۲). ساختار طرح‌ریزی بر خط بین دو عامل با ارتباطات محدود (o مشاهده دریافتی از محیط و a اقدام عامل) [۱۳].

۳- یافتن سیاست بهینه

هدف اصلی هر الگوریتم، طرح‌ریزی یافتن سیاست‌های بهینه‌ای است که اقدامات درست را اتخاذ کنند. هر عامل i یک سیاست محلی دارد که نگاشتی از پیشینه‌ها به اقدامات است $\delta i(hi) : Hi \rightarrow Ai$

و $\delta i(hi)$ نشان‌دهنده اقدامی است که به پیشینه hi منسوب می‌شود. سیاست‌های همه عامل‌ها با هم تشکیل یک سیاست مشترک را می‌دهد؛ در واقع یک سیاست مشترک مجموعه‌ای از سیاست‌های محلی است: $\delta=<\delta 1,\delta 2, \ldots, \delta n>$ برای هر عامل یکی و $\delta(h)$ نشان‌دهنده اقدام مشترک متعلق به پیشینه مشترک h است. هر عامل به‌طور مستقل همان طرح $\delta(h)$ را برای تیم محاسبه می‌کند و سپس بر اساس پیشینه محلی خود، بخشی از طرح را که مربوط به او است، اجرا می‌کند.

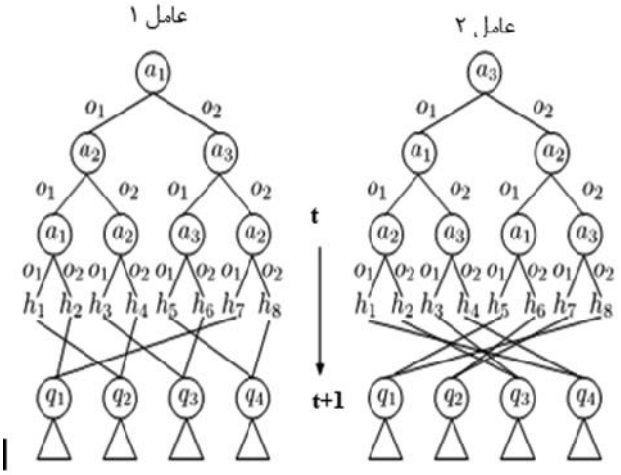
به‌منظور یافتن سیاست qi برای پیشینه hi عامل i، عامل‌ها باید درباره تمام پیشینه‌های ممکن h-I که به وسیله دیگران نگاه داشته شده و همچنین تمام سیاست‌های ممکن مربوط به آن‌ها استدلال کنند. به عبارت دیگر، ما نیاز داریم که یک سیاست مشترک δ بیابیم که تابع ارزش زیر را بیشینه کند:

$$V(\delta) = \sum_{h\in H}\sum_{s\in S} P(s|h)V(\delta(h),s) \qquad (۲)$$

1- Observe Orient Decide Act

در رابطه فوق، sحالت فعلی عامل در محیط، سیاست مشترک عامل‌ها با توجه به پیشینه h است. یافتن سیاست‌های غیرمتمرکز شبیه به ساخت درخت سیاست است که در الگوریتم‌های طرح‌ریزی DEC-POMDP به کار می‌رفت. همان‌طور که در شکل (۳) نشان داده شده است، پیشینه (h1, h2, . . ., hn) مسیرهایی از درخت از سمت ریشه به شاخه جاری هستند.

راه مستقیم یافتن بهترین سیاست مشترک این است که تمام نگاشت‌ها از پیشینه‌ها به زیر سیاست‌ها به شمار آیند و سپس با عملیات جستجو، بهترین آن‌ها انتخاب شود. در واقع، این مسئله با مسئله تصمیم‌گیری غیرمتمرکز که در مرجع [۴] آمده است، هم‌ارز است که ثابت شده است که پیچیدگی آن NP-HARD است. بنابراین در الگوریتم ارائه‌شده، ما با استفاده از توابع اکتشافی یک راه‌حل تقریبی می‌یابیم.



شکل (۳). یافتن سیاست‌ها برای طرح‌ریزی برخط را می‌توان به‌صورت پیمایش درخت سیاست نشان داد [۲].

این راه‌حل عبارت است از حل مسئله به‌عنوان یک برنامه خطی. مقدار سیاست مشترک یعنی π به‌صورت زیر محاسبه می‌شود [۲]:

$$V(\pi)=\sum_{h\in H}p(h)\sum_{\vec{q}}\prod_{i\in I}\pi_i(q_{i|h_i})Q(\vec{q},b(h)) \qquad (۳)$$

که p(h) توزیع احتمال پیشینه h است، b(h) حالت باوری است که توسط h استنتاج شده است و $Q(\vec{q},b(h))$ مقدار سیاست q در b(h) است. q سیاستی از زمان جاری به انتهای مسئله است، بنابراین $Q(\vec{q},b(h))$ مقداری است که عامل‌ها در گام‌های آتی، هنگامی‌که با یک توزیع حالت b (h) شروع می‌کنند، به آن دست می‌یابند که می‌توان آن را با یک تابع اکتشافی قابل قبول تخمین زد.

۳-۱- به‌کارگیری تابع اکتشافی پیشنهادی برای یافتن سیاست‌های تقریباً بهینه

همان‌طور که اشاره شد، یافتن مقدار بهینه سیاست‌ها دشوار است. چون برای حل آن باید به اندازه حل تمام DEC-POMDP کار کرد. مقدار بهینه می‌تواند با تابع اکتشافی معینی تخمین زده شود. در حالت ایده‌آل، تابع اکتشافی نه‌تنها باید مقدار ناگهانی اقدام مشترک را ارائه دهد، بلکه مقدارهای آتی نیز مورد انتظار هستند. رویکردی که در روش‌های موجود استفاده شده است، به‌کارگیری تابع اکتشافی MDP است. معمولاً، اکتشافی که نزدیک به مقدار بهینه باشد، ممکن است زمان بیشتری برای محاسبه صرف کند اما سبب نتیجه بهتری می‌شود و برعکس. از آنجا که در شرایط پیچیده مقدار نزدیک به بهینه به صرف زمان بیشتر ارجح است، ما هیورستیک POMDP را برای تخمین گام‌های آتی به کار می‌بریم.

$$Q(\vec{a},b)=\sum_{s\in S}b(s)\left[R(s,\vec{a})+\sum_{s'\in S}P(s'|s,\vec{a})\sum_{\vec{o}\in\Omega}O(\vec{o}|s',\vec{a})\,V_{POMDP}(b_{\vec{a}}^{\vec{o}})\right] \qquad (۴)$$

که $b_{\vec{a}}^{\vec{o}}$ حالت باور جانشین b با اقدام و مشاهده $\vec{a},\vec{o}$ است، S' حالت بعد از حالت s که احتمالاً عامل لحظه‌ای بعد به آن حالت خواهد رفت و V_{POMDP} تابع ارزش POMDP مورد نظر است [۶]. به‌طور شهودی، بکارگیری این تابع اکتشافی به آن معنا است که عامل‌ها مشاهدات خود را در هر گام بعدی با هم به اشتراک می‌گذارند.

$$V_{POMDP}=R(s)+\gamma\max(a')\sum p(s'|s,a)V(s') \qquad (۵)$$

رابطه فوق برای طرح‌ریزی یک عامل در محیط غیر قطعی به‌کار می‌رود [۶]. در این رابطه $V(s')$ تابع ارزش مشترک بدست‌آمده در حالت آتی عامل‌ها است و ضریب γ مقداری در بازه [۰،۱] است که برای تنظیم تأثیر میزان ارزش حالت‌های آتی نسبت به حالت‌های قبلی به کار می‌رود.

۳-۲- روش پیشنهادی برای یافتن دقیق‌تر سیاست‌ها: اتخاذ پیش‌بینی چند گام در درخت سیاست

برای پیش‌بینی تک‌گام، $V(s')$ را برابر $R(s')$ (بهره دریافتی از محیط در حالت) قرار می‌دهیم و برای پیش‌بینی تا چند گام فراتر و

یافتن دقیق‌تر سیاست را توسعه می‌دهیم. کارهای انجام‌شده، پیش‌بینی تک‌گامی را برای محاسبه مقدار گام‌های آینده به کار برده‌اند. بدین معنی که درخت سیاستی با فقط یک گره اقدام در نظر می‌گرفتند. این امر منجر به کاهش دست‌یابی به سیاست‌های دقیق و اتخاذ تصمیمات نادرست می‌شد. ما پیش‌بینی دو گام را به‌کار می‌بریم تا سیاست دقیق‌تری در طرح ریزی اتخاذ شود. بدین معنی که درخت سیاست نشان‌داده‌شده در شکل (۲) را تا عمق دوم توسعه می‌دهیم.

$$V(s') = R(s') + \gamma \max_{a'} \sum_{s''} P(s''|s',a')V(s'') . \qquad (۶)$$

در رابطه فوق، s حالت فعلی عامل در محیط، s' حالت بعد از حالت s که احتمالاً عامل لحظه‌ای بعد به آن حالت خواهد رفت و s'' دو حالت آتی پس از حالت کنونی یعنی s است. برای شروع فرآیند جستجو، هر سیاست محلی π با انتخاب یک اقدام تصادفی (ai) با توزیع یکنواخت مقداردهی می‌شود تا معین شود. سپس با استفاده از برنامه‌نویسی خطی[1]، بهترین سیاست ممکن که بهترین اقدامات را به دست دهد، می‌یابیم. به این صورت که هر عامل به‌صورت دوره‌ای انتخاب می‌شود و در حالی که سیاست دیگر عامل‌ها ثابت در نظر گرفته می‌شود، سیاست آن عامل بهبود داده می‌شود. برای تحقق این امر برای عامل iام بهترین اقدام ai* (و سیاست مربوط به آن qi) را در مجموعه hiها می‌یابیم که $V(\pi)$ در فرمول شماره (۳) بیشینه شود [۱۵].

۴- نتایج آزمایشات

برای اثبات تمامیت الگوریتم ارائه‌شده، الگوریتم پیشنهادی را روی مسئله مهم GRID SOCCER که از جمله جدیدترین سناریوهای ارائه‌شده برای آزمایش الگوریتم‌های DEC-POMDP است، آزمایش کردیم. البته مسائل آزمایشی ساده دیگری نیز برای طرح‌ریزی مشارکتی وجود دارد که اکثراً برای محیط‌های ایستا طراحی شده‌اند. از جمله این مسائل می‌توان Meeting in a Grid [9] و Cooperative Box-Pushing [10] را نام برد. ساختار این مسائل آزمایشی برای ابزار MADP طراحی شده است. ابزار MADP برای طرح‌ریزی مشارکتی مبتنی بر مارکوف در سیستم عامل لینوکس نوشته شده است [۱۱]. ابزار MADP محدودیت‌های زیادی دارد. از این رو ما از این ابزار استفاده نکرده و برای عمومیت بخشیدن به کاربرد کار خود، در محیط جاوا پیاده‌سازی‌های لازم را انجام دادیم.

سناریوی Grid Soccer نسبت به سناریوهای مذکور، علاوه‌بر پویایی بالای محیط آن، از تمامیت و سطح پیچیدگی بیشتری برخوردار است به طوری که تعداد حالات سناریوی آن بیش از ۲۰ برابر سایر سناریوها است و می‌توان گفت تنها بنچ مارک آزموده‌شده برای مدل‌های مارکوف نیمه‌رؤیت‌پذیر غیرمتمرکز برخط است. از این رو این سناریو را محور ارزیابی رویکرد ارائه‌شده قرار دادیم. در این سناریو دو عامل همکار در مقابل یک عامل حریف در یک فضای ۳×۳ و ۳×۲ به بازی فوتبال می‌پردازند. تعداد حالات فضای حالت در مسئله ۳×۳ برابر ۱۶۱۳۱ و در مسئله ۳×۲ برابر ۳۸۴۳ حالت متفاوت است.

روش ارائه‌شده در جاوا با پردازنده ۲٫۷ گیگاهرتز و حافظه RAM به میزان ۴GB پیاده‌سازی شده است. حالت اولیه بازی دو عامل در شکل (۴) نشان داده شده است که به طور تصادفی یا دلخواه انتخاب می‌شود و حالت موفقیت عامل‌ها پس از هشت گام زمانی در شکل (۵) نشان داده شده است که عامل‌ها به صورت مشارکتی توپ را وارد دروازه می‌کنند. توپ با رنگ زرد، عامل‌های همکار به رنگ قرمز و عامل حریف به رنگ آبی نشان داده شده است.

در جدول شماره (۱)، کارایی دو الگوریتم با ارتباطات محدود بر روی دو دامنه (فوتبال ۲×۳ و ۳×۳) مورد بررسی قرار گرفته است و بهره تجمعی میانگین (بهره)، میانگین زمان اجرای برخط در هر مرحله (زمان ((s)) و میانگین درصد ارتباط بین عامل‌ها در طول مدت طرح‌ریزی (ارتباط (٪)) با هر دو دامنه با روش ارائه‌شده در مقایسه با بهترین روش پیشین یعنی MAOP-COMM ارائه شده است.

تأثیر ارتباط بر میزان کارایی هر دو الگوریتم طرح‌ریزی در جدول شماره (۲)، برای سناریو فوتبال ۲×۳ و در جدول شماره (۳)، برای سناریوی ۳×۳ نشان داده شده است. همان‌طور که در جدول‌ها مشهود است، وقتی عامل‌ها هیچ ارتباطی با هم ندارند باز هم بهره دریافتی از محیط مطلوب است و این امر نشان می‌دهد که عامل‌ها در شرایطی که کانال ارتباطی قطع باشد نیز به‌خوبی طرح‌ریزی می‌کنند. البته همین الگوریتم در حالتی که ارتباطات آزاد و دائمی برقرار است، کارایی بیشتری از حالت بدون ارتباط دارد. چرا که عامل‌ها مشاهدات خود را دائماً با هم به اشتراک می‌گذارند و میدان دید محدود نیست. اختلاف بهره تجمعی در رویکرد با ارتباط و بدون ارتباط در سناریوی ۳×۳ بیشتر است. به نظر می‌رسد در سناریوهای بزرگ‌تر تأثیر ارتباطات آزاد بیشتر باشد؛ به بیان دیگر،

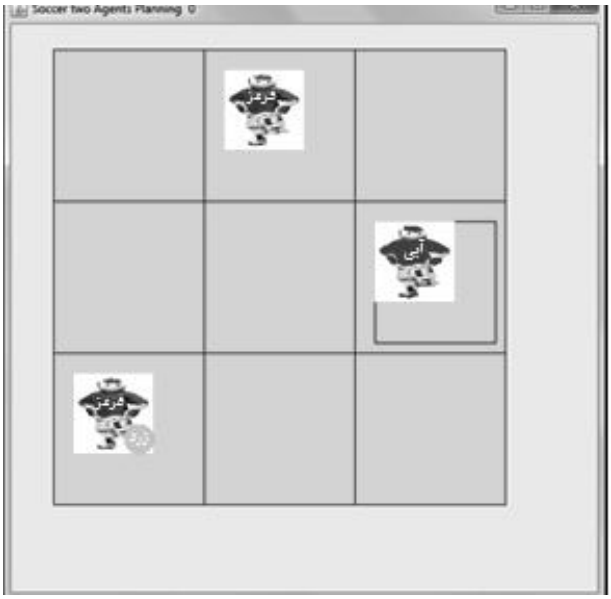
1- Linear Programming

هرچه سناریو بزرگ‌تر باشد، ارتباطات بیشتر، موجب می‌شود که عامل‌ها مشاهدات کامل‌تری داشته باشند و در نتیجه، بازده بیشتری در طرح‌ریزی به دست آورند.

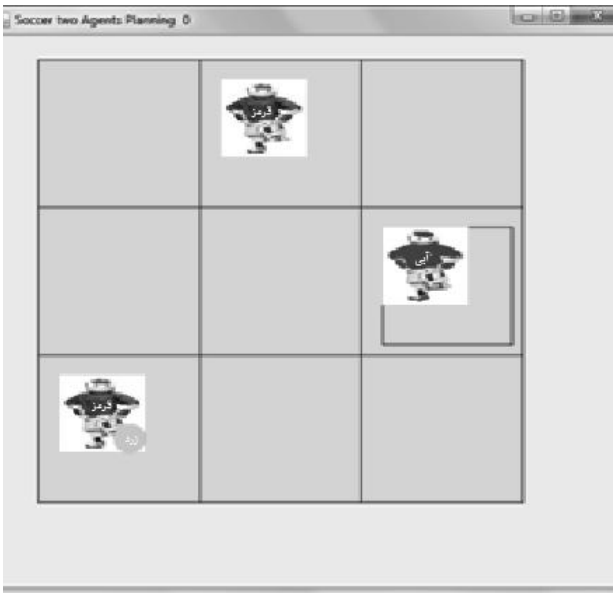
همان‌طور که در جدول‌ها قابل مشاهده است، مدت زمان پاسخگویی روش ارائه‌شده، از روش پیشین بیشتر است؛ اگر چه هر دو روش موجود در جدول دارای پیچیدگی زمانی NP-HARD هستند، علت افزایش زمان پاسخگویی در الگوریتم ما استفاده از پیش‌بینی دو گام بجای پیش‌بینی تک‌گام است؛ چرا که برای هر حالت ممکن از فضای مسئله درخت،سیاست را تا عمق دوم توسعه می‌دهد.

اما روش MAOP پیشین پیش‌بینی تک‌گام را برای یافتن سیاست اتخاذ کرده است. به علاوه تابع اکتشافی POMDP نیز به دلیل به‌کارگیری توزیع احتمال حالات بعدی ممکن و محاسبه رابطه پیچیده محاسبه باور برای حالات آتی به مقدار زیادی از سرعت اجرا می‌کاهد. با این وجود، میزان بهره دریافتی (آن که نشانگر کارایی طرح ریزی انجام شده است) به نحو چشمگیری افزایش یافته است.

از این رو، در شرایط بحرانی محیط‌های پیچیده که نیاز به تصمیم‌گیری دقیق، ولو با صرف زمان بیشتر، اهمیت فراوانی دارد، الگوریتم پیشنهادی نسبت به رویکرد پیشین در حوزه مدل‌های مارکوف، ارجح است.



شکل (۴). طرح ریزی در سناریوی Grid Soccer با دو عامل همکار با دامنه ۳×۳ (شروع بازی- توپ با رنگ زرد نشان داده شده است.)



شکل (۵). نتیجه اجرای الگوریتم طرح‌ریزی پیشنهادی در سناریوی Grid Soccer (رسیدن به هدف پس از ۸گام)

جدول (۱). مقایسه کارایی روش ارائه‌شده و MAOP-COMM

الگوریتم	بهره	مدت زمان(S)	ارتباط (٪)	تعداد دفعات
MAOP-COMM ۳×۳	۲۹۶	۲،۳	۲۷	۲۰
	۱۶۷۹،۵	۲،۵	۲۶،۹	۱۰۰
روش پیشنهادی ۳×۳	۷۱۰،۸	۱۰	۲۳	۲۰
	۳۵۵۸،۲	۳۱	۲۴	۱۰۰
MAOP-COMM ۲×۳	۲۹۰،۶	۰،۲۸	۱۴،۸	۲۰
	۱۹۳۳،۹	۰،۱۶	۱۵،۴	۱۰۰
روش پیشنهادی ۲×۳	۱۴۲۴،۳۴	۷	۱۰،۴	۲۰
	۳۶۶۸،۴	۲۹	۱۲	۱۰۰

جدول (۲). مقایسه تأثیر ارتباط در روش ارائه‌شده و MAOP (فوتبال ۲×۲ تعداد حالات ۳۸۴۳، تعداد مشاهدات :۱۱، تعداد اقدامات: ۶)

الگوریتم	بهره	مدت زمان(s)	ارتباط (٪)	تعداد دفعات
MAOP	۱۸۰،۵	۰،۲۵	۰	۲۰
	۱۱۵۷،۸	۰،۱۴	۰	۱۰۰
روش پیشنهادی- بدون ارتباط	۶۷۰،۲	۱۰	۰	۲۰
	۳۱۶۱،۲	۳۰	۰	۱۰۰
MAOP FULL-COMM	۳۷۳،۹	۰،۰۱	۱۰۰	۲۰
	۱۹۳۳،۶	۰،۰۱	۱۰۰	۱۰۰
روش پیشنهادی- ارتباطات آزاد	۷۰۹،۲	۹	۱۰۰	۲۰
	۳۵۵۹،۶	۳۱	۱۰۰	۱۰۰

با توجه به جامعیت الگوریتم ارائه‌شده در پوشش پارامترهای محیط‌های پیچیده، می‌توان از آن برای طرح‌ریزی در حوزه هایی چون یگان‌های رزمی توپخانه نیز استفاده کرد. برای این امر کافی است جزئیات عملیاتی سناریوی مورد نظر در الگوریتم گنجانده شود. مثلاً در طرح‌ریزی برای عامل‌های توپخانه به جای ارتباطات محدود باید ارتباط آزاد باشد. چون در میدان نبرد، عامل‌های توپخانه باید به‌طور دائم از وضعیت عامل‌های همکار مطلع شوند.

۵- نتیجه

در این مقاله، با ارتقاء الگوریتم اکتشافی تابع ارزش و به‌کارگیری پیش‌بینی دو مرحله‌ای در استراتژی یافتن سیاست بهینه و اخذ تصمیم صحیح، رویکرد MAOP-COMM (از جمله جامع‌ترین الگوریتم طرح‌ریزی برخط موجود) را بهبود دادیم تا الگوریتم حاصل علاوه بر اینکه پارامترهای محیط‌های پیچیده را در خود لحاظ می‌کند، از کارایی مضاعفی نیز برخوردار شده و طرح‌ریزی بسیار دقیق‌تری نسبت به الگوریتم‌های اخیر انجام دهد. نتایج آزمایشات نشان می‌دهد اگرچه زمان طرح ریزی الگوریتم جدید نسبت به الگوریتم‌های قبلی به‌دلیل به‌کارگیری پیش‌بینی دوگام (که برای هر

جدول (۳). مقایسه تأثیر ارتباط در روش ارائه‌شده و MAOP (برای فوتبال ۳×۳ تعداد حالات ۱۶۱۳۱، تعداد مشاهدات :۱۱، تعداد اقدامات: ۶)

الگوریتم	بهره	مدت	ارتباط	تعداد
MAOP	۱۹۰،۷	۱،۹	۰	۲۰
	۸۰۳،۶	۱،۹۶	۰	۱۰۰
روش پیشنهادی- بدون ارتباط	۶۷۰	۱۰	۰	۲۰
	۲۹۵۳،۳	۳۲	۰	۱۰۰
MAOP FULL-COMM	۳۵۶	۰.۰۱	۱۰۰	۲۰
	۱۸۰۸،۲۰	۰.۰۱	۱۰۰	۱۰۰
روش پیشنهادی- ارتباطات	۷۱۹،۲	۹.۵	۱۰۰	۲۰
	۳۵۴۹،۲	۳۲	۱۰۰	۱۰۰

حالت، درختی به عمق دو از حالات ممکن آتی را توسعه می‌دهد) بیشتر است، اما میانگین بهره تجمعی حاصل از طرح‌ریزی بسیار بهتر است. از این رو، در شرایط بحرانی محیط‌های پیچیده که نیاز به تصمیم‌گیری دقیق، ولو با صرف زمان بیشتر، اهمیت فراوانی دارد، الگوریتم پیشنهادی نسبت به رویکرد پیشین در حوزه مدل‌های مارکوف برای شرایط پیچیده ارجح است. در کارهای آتی می‌توان با هرس کردن درخت سیاست و حذف حالات غیرمفید، پیچیدگی زمانی الگوریتم را نیز کاهش داد.

۶- مراجع

[1] M. Weerdt and A. Mors, "Multi-agent Planning an introduction to planning and coordination," Dept. of Software Technology, Delft University of Technology, 2005.

[2] F. Wu, S. Zilbersteinand, and X. Chen, "Online Planning with bounded Communication," Artificial Intelligence Journal, vol. 175, pp. 487–511, 2011.

[3] C. Amato, J. S. Dibangoye, and S. Zilberstein,

"Incremental policy generation for finit horizon DEC-POMDPs," Proceedings of the 19th International Conference on Automated Planning and Scheduling, pp. 2–9, 2009.

[4] J. Tsitsiklis and M. Athans, "On the complexity of decentralized decision making and detection problems," IEEE Transaction on Automatic Control vol. 30, pp. 440–446, 1985.

[5] S. Kruk, "Planning with Multiple Agents," Seminar on Autonomous Learning System, 2013.

[6] S. Seuken and S. Zilberstein, "Formal models and algorithms for decentralized decision making under uncertainty," Springer Science Conference on Auton Agent Multi-Agent Systems, 2008.

[7] D. Alberts and R. Hayes, "Planning Complex Endeavors," 1th Edition, CCRP Publication, Washington D.C, 2007.

[8] E. Shahbazian , D. Blodgett, and P. Labbe, "The Extended OODA Model for Data Fusion Systems," Proceedings of 4th International Conference on Information Fusion, Montreal Canada, 2001.

[9] C. Amato, D. Bernstein, and S. Zilberstein, "Optimizing fixed-size stochastic controllers for POMDPs and decentralized POMDPs," SPRING COMP. SCI., LLC, 2009.

[10] S. Seuken and S. Zilberstein, "Improved Memory-Bounded Dynamic Programming for Decentralized POMDPs," Proc. of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, 2012.

[11] F. Oliehoek and M. Spaan, "MADP Toolbox 0.2," Technical Report, Informatics Institute, Amsterdam University, 2009.

[12] B. Clement and K. Decker, "Multiagent Planning: A Survey of Research and Applications," Seventh Pacific Rim International Workshop on Multi-Agents (PRIMA), 2004.

[13] S. Sadati, M. NaghianFesharaki, and M. H. Momeni Azandaryani, "Online Collaborative Planning in Command and Control Domain," 7'Th Conference on Command and Control (C4I), Iran, 2013.(In Persian)

[14] S. Sadati, M. NaghianFesharaki, and S. M. Hoseini, "Survey of Planning Parameters for Command and Control Environments in Existing Planning Models," 7'Th Conference on Command and Control (C4I), Iran, 2013. (In Persian)

[15] S. Sadati, M. NaghianFesharaki, and M. H. Momeni Azandaryani, "Online Collaborative Planning in Command and Control Domain," 7'Th Conference on Command and Control (C4I), Iran, 2013.(In Persian)